ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

PRODUCTION NOTE

University of Illinois at
Urbana-Champaign Library
Large-scale Digitization Project, 2007.

LIBRARY TRENDS
VOLUME 48, NUMBER 2, 283-524
FALL 1999

Progress in Visual Information Access and Retrieval

Beth Sandore

Issue Editor

UNIVERSITY OF ILLINOIS
GRADUATE SCHOOL OF
LIBRARY AND INFORMATION SCIENCE

LIBRARY TRENDS

Library Trends, a quarterly thematic journal, focuses on current trends in all areas of library practice. Each issue addresses a single theme in depth, exploring topics of interest primarily to practicing librarians and information scientists and secondarily to educators and students.

Library Trends is published four times annually—in summer, fall, winter, and spring—by the Graduate School of Library and Information Science at the University of Illinois, Urbana-Champaign, 501 E. Daniel Street, Champaign, IL 61820-6211.

Subscriptions: Institutional rate is $85 per volume (plus $7 for overseas subscribers). Subscriptions for an individual are $60 (plus $7 for overseas subscribers). Registered students may subscribe for $25 (plus $7 for overseas subscribers). Individual issues are $18.50 (shipping included); back issues other than those from the present year are $10 (plus shipping). Claims for missing numbers should be made within six months following the date of publication. All foreign subscriptions and orders must be accompanied by payment.

Address orders to: University of Illinois Press, Journals Department, 1325 S. Oak Street, Champaign, IL 61820. For out-of-print issues, contact Bell & Howell Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346. **Postmaster:** Send change of address to University of Illinois Press, 1325 S. Oak Street, Champaign, IL 61820-6903.

 ISSN 0024-2594.
Postage paid at Champaign, Illinois.

This journal is abstracted or indexed in *Library and Information Science Abstracts, Current Contents, Current Index to Journals in Education, Information Science Abstracts, Library Literature, PAIS,* and *Social Sciences Citation Index.*

Procedures for Proposing and Guest Editing an Issue of Library Trends

We encourage our readers to submit ideas for future *Library Trends* themes; issue topics are developed through recommendations from members of the Publications Committee and from reader suggestions. We also encourage readers to volunteer to be issue editors or to suggest others who may be willing to be issue editors.

The style and tone of the journal is formal rather than journalistic or popular. *Library Trends* reviews the literature, summarizes current practice and thinking, and evaluates new directions in library practice. Papers must represent original work. Extensive updates of previously published papers are acceptable, but revisions or adaptations of published work are not sought.

An issue editor proposes the theme and scope of a new issue, draws up a list of prospective authors and article topics, and provides short annotations of the article's scope or else gives a statement of philosophy guiding the issue's development. Please send your ideas or inquiries to F. W. Lancaster, Editor, Publications Office, 501 E. Daniel Street, Champaign, IL 61820-6211.

LIBRARY TRENDS

Fall 1999 48(2) 283-524

Progress in Visual Information Access and Retrieval

Beth Sandore

Issue Editor

UNIVERSITY OF ILLINOIS
GRADUATE SCHOOL OF
LIBRARY AND INFORMATION SCIENCE

Progress in Visual Information Access and Retrieval

CONTENTS

Introduction

BETH SANDORE

THE DEVELOPMENT OF INNOVATIVE RETRIEVAL APPROACHES for access to visual information is among the most significant of technological, conceptual, and institutional challenges for the library and information science community. On a technological level, Gupta and Jain (1997) suggest that visual information retrieval involves a hybrid of older computer science disciplines, including the analytical component of computer vision and the query component of database systems. On a conceptual level, it is clear that humans employ a variety of socio-cognitive processes, as well as sensory skills, in the retrieval and evaluation of visual information. From an institutional standpoint, libraries, archives, and museums have entered into digitization projects, varying in scope and scale, the results of which are beginning to challenge the manner in which information is described, stored, and delivered. While visual resources have long been part of the slide library that supports use in the arts and humanities communities, Charles Rhyne (1996) indicates that technology has enabled a rapid increase in the use of pictures in other disciplines as well as by the general public (p. 4).

The primary goal of this issue of *Library Trends* is to present practitioners, researchers, and educators in the areas of library and information science, archives, and museums, as well as "imagists" working with visual resources in any setting, with a current perspective on the development of visual information retrieval and access tools. The issue's scope is limited to the analysis and retrieval of bit-mapped or raster images and video (images that are comprised of pixels of varying color information values) and does not include work with vector graphics (images encoded as numeric formulas that represent lines and curves—e.g., Geographic

Beth Sandore, Digital Imaging Initiative, 452 Grainger Engineering Library, University of Illinois, Urbana, IL 61801
LIBRARY TRENDS, Vol. 48, No. 2, Fall 1999, pp. 283-288

Information Systems [GIS]). The contributions provide perspectives from researchers and practitioners—specialists in the areas of library and information science and computer science. In planning this issue, a conscious effort was made to include a perspective on the developing foundation of visual information retrieval, as well as work representing current and experimental systems. The issue is divided into three sections—I. Foundations of Intellectual Access to Visual Information, II. Implementation and Evaluation, and III. Experimentation.

Since 1988, two issues of *Library Trends* have been devoted to various aspects of image and multimedia information retrieval. In each issue, the editors call for a synergy across the disciplines that develop image retrieval systems and those that utilize these systems. Stam and Giral (1988), in the issue of *Library Trends* titled "Linking Art Objects and Art Information," emphasize the need for a thorough understanding of the visual information-seeking behaviors of image database users. Writing in a 1990 issue of *Library Trends* devoted to graphical information retrieval, Mark Rorvig (1990) takes up the fundamental issue that "what can be listed cannot always be found" and uses that statement as a framework for examining progress in intellectual access to visual information. In the ensuing decade, several critical events have unfolded that have brought about some of the needed collaboration across disciplines and have enhanced the potential for advancements in the area of visual information retrieval.

First, the field of computer vision has grown exponentially within the past decade, producing tools that enable the retrieval of visual information, especially for objects with no accompanying structural, administrative, or descriptive text information. Second, the Internet, more specifically the Web, has become a common channel for the transmission of graphical information, thus moving visual information retrieval rapidly from stand-alone workstations and databases into a networked environment. Third, the use of the Web to provide access to the search and retrieval mechanisms for visual and other forms of information has spawned the development of emerging standards for metadata about these objects as well as the creation of commonly employed methods to achieve interoperability across the searching of visual, textual, and other multimedia repositories. Practicality has begun to dictate that the indexing of huge collections of images by hand is a task that is both labor intensive and expensive—in many cases more than can be afforded to provide some method of intellectual access to digital image collections. In the world of text retrieval, text "speaks for itself" whereas image analysis requires a combination of high-level concept creation as well as the processing and interpretation of inherent visual features. In the area of intellectual access to visual information, the interplay between human and machine image indexing methods has begun to influence the development of visual information retrieval systems. Research and application by the visual

information retrieval (VIR) community suggests that the most fruitful approaches to VIR involve analysis of the type of information being sought, the domain in which it will be used, and systematic testing to identify optimal retrieval methods.

Section I—"Foundations of Access to Visual Information"—is intended to provide a background in the familiar concept-based approach to describing and retrieving images, as well as the more recently developed content-based approach to visual information retrieval using inherent features such as color, shape, and texture. The importance of the articles in this section cannot be over-emphasized. In their own way, each clarifies the inevitable need to consider the interaction between high-level semantic concepts and inherent content in VIR. Content retrieval, the area which is newest to the library and information science community, will demand increased understanding and analysis in order to determine its value to users as we build more robust and lasting visual information retrieval systems. The authors in section I emphasize the need for a greater understanding of the interplay between concept-based indexing (performed by humans) and the automatic or semi-automatic process of indexing an image or a video sequence (using software) based on inherent image attributes. In "Intellectual Access to Images," Hsin-liang Chen and Edie M. Rasmussen explore current image retrieval systems and analyze the methods that have been employed to provide intellectual access to the various image collections. Throughout the article, the authors focus on the problems that are inherent in image description and access with the objective of identifying traditional and new solutions to these challenges. The second contribution, by P. Bryan Heidorn, presents a framework for understanding image retrieval from the standpoint of the user's cognitive models for seeking visual information. Heidorn examines the process by which models, based on linguistic and inherent visual attributes, are constructed and employed by users in seeking visual information that answers particular queries. He also discusses the types of information that can be found to have common values in the socio-cognitive sense. In "Computer Vision Tools for Finding Images," David Forsyth describes and discusses the use of two types of methods that have been developed in the computer vision field to facilitate the searching of images by inherent content. Forsyth groups these methods into two categories—"appearance methods" that compare images based on their overall content (e.g., color histograms, texture histograms, spatial layout), and "finding methods" that focus on matching subparts of images with the goal of identifying and finding specific objects. Forsyth explains, in terms understandable to a broad range of readers, the complexities involved in identifying and matching inherent features within whole images and corresponding objects within segments of images. He is careful to explain that computer vision tools cannot be used in monolithic ways to resolve user queries of large

image collections but rather describes areas in which they have been found to be most useful and promising for future development.

Section II—"Implementation and Evaluation"—focuses more specifically on the implementation and evaluation of visual information retrieval systems with cultural heritage information since this is a primary interest of libraries, museums, and archives. In the first article in the section, Teresa Grose Beamsley addresses the challenge of securing and ensuring image integrity—an issue that is integral to the quality of VIR system search results yet often goes unremarked in discussions beyond the initial point of digital capture. Beamsley's article, "Securing Digital Image Assets," examines the issues involved in the process of securing the image content in a VIR system. Beamsley indicates that images delivered across the Web are usually low-quality compressed derivatives of higher quality archival digital images. Often, the only tenuous link between the low-quality derivative image and its original digital image is contained in the textual metadata that accompanies the image. Beamsley examines various approaches that institutions can use to secure the integrity of this representative information while pointing out the concomitant challenges in doing so. Throughout her article, Beamsley focuses on the need to achieve a balance between the desire for ownership and authenticity of images and the provision of open access to cultural heritage materials, particularly in public institutions.

In the second article in this section, "Getting the Picture," Caroline Arms provides a thorough description and analysis of Library of Congress efforts to provide access to visual information from their Prints and Photographs Division collections. The Library of Congress' work in this area is of international significance because they are one of the few institutions that makes images in their collections and experimental projects publicly accessible through the Web (except in cases of copyright and ownership restrictions), as well as information regarding the technical underpinnings of their efforts. Arms's account gives important insights into the institutional challenges of providing unified public access to disparate digital collections while addressing the special issues associated with VIR and other programmatic concerns. The article will be useful for any institution considering the development of a digital library or facing the organizational challenges of identifying unified aspects of collections that are otherwise disparate in physical location and diverse in content.

Christie Stephenson, in "Recent Developments in Cultural Heritage Image Databases," uses the Museum Educational Site Licensing (MESL) Project as a point of departure in her exploration of developments in the broad area of cultural heritage image databases. Stephenson draws on her own experience in the management of the MESL project as well as on the work of others to identify the various factors that are known to affect VIR including metadata quality, image quality, display and manipulation

features, and the diverse retrieval results due to a variety of methods employed in search engine indexing and retrieval. She also reviews examples of recent work in the development of federated image repositories at various institutions, as well as advances in user interface design for VIR. Stephenson's writing reiterates the point made by Forsyth, albeit from a different perspective, that it is critical to identify the information-seeking needs of specific user groups in order to tailor the most effective retrieval methods and interfaces to specific domains of use.

In her article "Evaluation of Image Retrieval Systems: The Role of User Feedback," Samantha Hastings reviews problems in current VIR evaluation research, presents the preliminary results of a Web-based study of user searching of an art image database, and proposes a framework for user-centered evaluation studies of VIR systems. In particular, Hastings's findings note that over half the user queries submitted in the study were satisfied by the review of thumbnail images. Further, Hastings emphasizes the importance for users to have the capability to construct customized browsing approaches to retrieved image sets and to manipulate images to view more visual detail and to compare more than one image.

Section III—"Experimental Approaches"—presents articles describing three research projects that examine various aspects of image or combined image and text retrieval methods. The articles by Rohini K. Srihari and Zhongfei Zhang and by Neil C. Rowe focus on experimental systems that employ both text and image analysis in the development of effective methods for retrieving highly relevant image sets. Srihari and Zhang define their subject domain specifically. They focus on analyzing faces in pictures and their related captions taken from Web-based newsfeeds such as MSNBC and CNN. From this standpoint, these authors are able to tailor their retrieval algorithms to achieve a fairly high level of precision for this domain of multimedia information. Rowe uses similar methods to analyze databases of images that feature a range of activities at a naval aircraft test facility. Rowe's approach differs from that of Srihari and Zhang in its objective, which he states is to broaden the applicability of linguistic and text processing routines in an attempt to create a more general retrieval process targeted toward increasing precision in the searching of Web-based information in general. The final article in this issue, by Yong Rui et al., focuses entirely on experimental methods using computer vision tools, testing various methods of image analysis (vector analysis, Boolean, fuzzy match) that are commonly used in text processing, and using relevance feedback to "train" the search engine and improve the relevance of the retrieved image result set. Rui et al. use an experimental multimedia system, MARS (Multimedia Analysis and Retrieval System), which they have developed over a period of several years, to test various approaches to the analysis and retrieval of inherent image features. They have tested their methods with various

image sets, including a set of cultural heritage images of artifacts from the UCLA Fowler Museum of Cultural History.

The work represented in this issue suggests that a number of professional communities are contributing different but essential components to the development of useful and innovative image retrieval systems. In spite of the great technology strides in multimedia, image database developers and image content holders continue to grapple with the fluid issues of organization, access, retrieval, delivery, and representation. The words of art historian Barbara Stafford (1996) describe the differential treatment accorded visual imagery over the centuries in Western culture, and they express the hope that computers will be the tool that enables imagery to become a trusted, valued, and rich vehicle (similar to text) for information delivery:

> Yet in spite of the arrival of what I have termed the "age of computerism"—rapidly replacing modernism and even postmodernism—a distorted hierarchy ranking the importance of reading above that of seeing remains anachronistically in place. All the while, computers are forcing the recognition that texts are not "higher" durable monuments to civilization compared to "lower" fleeting images. These marvelous machines may eventually rid us of the uninformed assumption that sensory messages are incompatible with reflection. (p. 4)

Despite Stafford's apparent ambivalence, the significant levels of traffic on the Web support the perspective that technology has begun to fuel an important shift in the value that society has previously placed on the written word over things visual. Computers now enable users to incorporate images of art and other works into their own personal information contexts—images which have for centuries been a powerful and efficient medium for conveying landmark concepts, emotions, and events. The concomitant challenge for libraries, museums, and archives also involves a shift—not only in technology and practice but also in focus—i.e., to equip ourselves with an effective understanding of the similarities and differences between text and multimedia information retrieval, and to use this knowledge as a foundation for developing effective access and archiving methods.

REFERENCES

Gupta, A., & Jain, R. (1997). Visual information retrieval. *Communications of the ACM, 40*(5), 70-79.

Rhyne, C. S. (1996). *Computer images for research, teaching, and publication in art history and related disciplines.* Washington, DC: The Commission on Preservation & Access.

Rorvig, M. E. (1990). Introduction. *Library Trends, 38*(4), 639-643.

Stafford, B. M. (1996). *Good looking: Essays on the virtue of images.* Cambridge, MA: MIT Press.

Stam, D. C., & Giral, A. (Eds.). (1988). Linking art objects and art information (theme issue). *Library Trends, 37*(2), 117-264.

Foundations of Access to Visual Information

Intellectual Access to Images

HSIN-LIANG CHEN AND EDIE M. RASMUSSEN

ABSTRACT

CONVENIENT IMAGE CAPTURE TECHNIQUES, inexpensive storage, and widely available dissemination methods have made digital images a convenient and easily available information format. This increased availability of images is accompanied by a need for solutions to the problems inherent in indexing them for retrieval. Unfortunately, to date, very little information has been available on why users search for images, how they intend to use them, as well as how they pose their queries, though this situation is being remedied as a body of research begins to accumulate. New image indexing methods are also being explored. Traditional concept-based indexing uses controlled vocabulary or natural language to express what an image is or what it is about. Newly developed content-based techniques rely on a pixel-level interpretation of the data content of the image. Concept-based indexing has the advantage of providing a higher-level analysis of the image content but is expensive to implement and suffers from a lack of interindexer consistency due to the subjective nature of image interpretation. Content-based indexing is relatively inexpensive to implement but provides a relatively low level of interpretation of the image except in fairly narrow and applied domains. To date, very little is known about the usefulness of the access provided by content-based systems, and more work needs to be done on user needs and satisfaction with these systems. An examination of a number of image database systems shows

Hsin-liang Chen, School of Information Sciences, University of Pittsburgh, 646 IS Building, 135 North Bellefield Avenue, Pittsburgh, PA 15260
Edie Rasmussen, School of Information Sciences, University of Pittsburgh, 646 IS Building, 135 North Bellefield Avenue, Pittsburgh, PA 15260
LIBRARY TRENDS, Vol. 48, No. 2, Fall 1999, pp. 291-302

the range of techniques that have been used to provide intellectual access to image collections.

Introduction

With the rapid development of computing technologies, particularly in storage, display, and telecommunications, access to digital images has become widespread. At the same time, the ease with which images can be incorporated into software packages for display, publication, and dissemination has increased the perceived information need of users for images. This greatly increased need for, and access to, images has focused attention on the problems inherent in image description, particularly from the perspective of image indexing and retrieval. Researchers in the fields of library and information science, computer science, medical informatics, cognitive science, and so on, have brought their different points of view to address the problems inherent in image indexing. The development and use of controlled vocabularies for image indexing has always been an area of interest, and the exploration of natural language for image description is an area of ongoing research. A relatively new research area, drawing on the pixel-level data that comprise digital images, is content-based retrieval, which automatically extracts index features such as color, texture, and shape from the image file. Other researchers are examining the potential of combined sources of evidence using natural language text, such as captions, to assist in the automatic interpretation of digital images. A welcome development in the study of image access is the focus of a number of researchers on questions underlying users' access to images—i.e., how images are perceived and described, what information needs exist, and how users of pictorial information determine what is useful to them. The answers to these questions will inform the design of a new generation of image retrieval systems that will better meet the needs of users by employing technologies in useful and creative ways.

The discussion will focus on the problems inherent in image description and access, with a perspective on traditional and new solutions. Recent developments in intellectual access to images will be surveyed and contrasted with software-based analysis of image content. A more detailed survey of this research is given by Rasmussen (1997). Lancaster (1998) extends the discussion of indexing to multimedia sources.

Challenges in Image Indexing

Images bring with them problems of description and access more complex than those of text. While text can be indexed manually, it can also be retrieved directly using, as access points, the natural language that it contains. While this retrieval is imperfect, it does provide a means of access independent of human indexing. Digital images are composed of pixels arranged in an infinite variety of patterns and, in general, it is im-

possible to predict the particular pattern that would match an information need. At present, only relatively low-level attributes of images can be queried directly (for instance, color and texture), and these attributes do not carry the meaning of the image with them.

Even where human indexing of the image is undertaken, it is difficult to reach agreement on the content and meaning of the image, or on what aspects are appropriate for indexing. The same image may mean different things to different people and may be used to project a different meaning at different times depending on the way it is used or the aspect that is the focus of attention or the context it is chosen to illustrate.

In general, it is easier to determine a picture's content than to interpret what it is about, and this distinction has engaged many scholars. Krause (1998) distinguishes between "hard" indexing (the description of what an indexer can see in the frame), and "soft" indexing ("aboutness," the image as stimulus). He says:

> We know that pictures provoke reaction, stimulate ideas, rekindle memories. They are powerful instruments in story telling, teaching, propaganda, and numerous other fields. Therefore, it is important that libraries provide access to images which illustrate ideas, even abstract ones like hunger, or the experience of hunger .... If we can index this aspect of the picture, we make it easily available to users requiring such an image; we make our collection more accessible. (pp. 73-74)

A number of authors (such as Shatford, 1986) have based their analysis of image indexing on the theories of the art historian Panofsky (1939), who identified three levels of meaning in works of art. At the first, or pre-iconographic, level, subject matter was designated as factual ("ofness") or expressional ("aboutness"), and based on the objects and events in an image as it could be interpreted through everyday experience. At the second, or iconographic, level, interpretation requires some cultural knowledge of themes and concepts (not "a sailor" but "Ulysses"). The third or iconological level requires interpretation at a sophisticated level using world and cultural knowledge plus a deeper understanding of the history and background of the work. Shatford (1986) suggests that this third level cannot be indexed with any degree of consistency. Svenonius (1994) points out that "indexing aboutness at the iconographic level is equally problematic" (p. 603), since what is symbolized is not always evident, nor is there always a simple referent to it.

Shatford (1986) uses Panofsky's levels of meaning to explore the kinds of subjects an image might have, proposing "Generic Of," "Specific Of," and "About" with facets answering the questions Who? What? When? and Where? Interestingly, a preliminary attempt in the Hulton study (described later in this article) to categorize queries posed to an image database according to Panofsky's levels of meaning was not successful, suggesting that

they did not translate well from the area of Renaissance art to a more general domain (Enser & McGregor, 1993).

Markey (1984) studied interindexer consistency by nonspecialists using a free vocabulary to index pictorial works of art, finding terminology consistency scores of 7 percent and concept consistency scores of 13 percent. While interindexer consistency has always been problematic, even in text, these figures do serve to illustrate the imperfect level of agreement in subject analysis of images. Clearly, image analysis can be carried out at many levels, from the primitive (What colors are present? What shapes?) to more analytical but general (What objects appear in the image? What is this a picture of?) to a more culturally dependent interpretation (What specific individual or thing is portrayed? What is the mood? What metaphor or lesson is presented?). Choosing an approach to image indexing may require a compromise based on what the system is capable of delivering and what the users of the system would like in an optimal retrieval environment. The question of user need for images is at present relatively little studied.

Studies of Users' Image Queries

Before considering how image access has been provided, it is worth considering what we know about users' information needs and how users present queries to image databases. Probably the most extensive study to address this question is the "Hulton Study," Enser and McGregor's (1993) examination of the queries addressed to the Hulton Deutsch picture collection. The Hulton Deutsch Collection is a major picture archive of news and current affairs, historical landscapes and portraits, and other collections used primarily by the press. Enser and McGregor examined 2,722 requests and found that they could be mapped into four categories along two dimensions: unique ("Kenilworth Castle") or non-unique ("dinosaurs") and refined (e.g., specified by activity, time period, and so on) or nonrefined. An example of a query in the unique refined category is "Edward VIII looking stupid" and in the nonunique refined category is "couples dancing the Charleston." Interestingly, only requests for unique unrefined subjects were easily satisfied by the Gibbs-Smith classification scheme being used by the picture archive (Enser, 1995). The Hulton Study was subsequently extended to seven additional picture libraries/archives in the United Kingdom, five of which were concerned with still images (Armitage & Enser, 1997). They arrived at a mode and facet analysis adopted from Panofsky (1939) and refined by Shatford (1986).

In a smaller-scale study, Hastings (1995) examined the queries of a specific user group—i.e., art historians—to a collection of Caribbean art images. She identified queries at four levels of complexity, ranging from simple level one queries for who, what, where to level four queries for meaning, subject, and why? Some of the simpler queries could be an-

swered without images while, at the most complex level, text and image alone was sometimes not sufficient to answer the queries.

Another interesting study reviewed queries presented to NLM's Prints and Photographs Collection. Keister (1994) found that descriptions of concrete image elements made up a significant proportion of picture requests, and these elements were worth cataloging in some detail. However, she cites examples in which images are described in terms of the visual message of the picture—e.g., a "warm picture of a nurse, mother, and baby" (p. 10). Word-images based on a particular communication need arose frequently, and users often described and used images in ways different from their original intent.

While there begins to be a body of research addressing the question of image information needs, the studies are fragmented. The Hulton Study remains the only study of its scale to examine information needs in a nondomain-specific environment.

Image Attributes

There is as yet no general agreement on what attributes of an image should be indexed. Shatford (1986) indicates that it is much easier to index an image for a collection with some specific use than one for use by a heterogeneous group. In the latter case, the subject orientation of users and the information need that will lead them to pose queries to the collection cannot be anticipated, and hence the dimensions along which the collection should be indexed cannot be predicted.

Research by Jorgensen (1998), in which participants were asked to write descriptions of color images, suggested four perceptual classes as a minimal framework for image indexing: objects (the largest set in her study), people, color, and location. Content/story attributes were also identified as significant for image description. Jorgensen (1998) points out the need to include interpretive as well as perceptual attributes, a conclusion supported by her previous research (1995). She indicates that "the disjunction between these results and those attributes typically addressed in traditional image indexing systems suggest revisiting assumptions upon which image indexing and retrieval systems are being created" (p. 172).

In order to determine what image attributes should be used to provide access, Layne (1994) proposes four categories: (1) "biographical" attributes that deal with the images' origin and provenance; (2) subject attributes (the "most problematic and least objective" [p. 584]); (3) exemplified attributes that seem to be physical characteristics such as medium, and (4) relationship attributes (relationship to other images or texts). It is the subject attributes which are addressed here; the two main approaches to image indexing, concept-based and content-based, differ in the level of interpretation that they bring to the indexing process and will be discussed separately.

Concept-Based Image Indexing and Retrieval

Concept-based retrieval refers to retrieval from text-based indexing of images, which may use a controlled vocabulary or natural language text or captions, and range from the purely descriptive ("Winston Churchill," "a duck on a pond") to the abstract or subjective ("poverty," "despair").

Controlled vocabularies have been developed for use in specific collections such as Western art or images of historical costume. A particularly ambitious undertaking is ICONCLASS, an early classification system for Western art developed in the 1940s by van de Waal at the University of Leiden. Nine areas are covered: (1) Religion, Magic and the Supernatural; (2) Nature; (3) Man (as a biological entity); (4) Society, Civilization, and Culture; (5) Abstract Concepts; (6) History; (7) The Bible; (8) Non-Classical Myths, Tales and Legends; and (9) Classical Mythology and History (Sherman, 1987). These subjects are subdivided using an alphanumeric notation covered in seven volumes of subject headings. ICONCLASS has been used for DIAL (Decimal Index to the Art of the Lowlands), the Marburger Index to works of art in Germany, van Straten's index of Italian Prints, and American paintings in the Courtauld Institute (Roberts, 1988).

Two controlled vocabularies that were developed relatively recently are the *Art & Architecture Thesaurus* (*AAT*) (Oxford University Press, 1990) and the Library of Congress *Thesaurus for Graphic Materials* (Library of Congress, 1995). The *Art & Architecture Thesaurus (AAT)* covers the history and making of the visual arts and is geographically and historically comprehensive but lacks coverage of iconographical themes (Petersen, 1990). The vocabulary of nearly 120,000 terms is structured under seven facets (e.g., physical attributes, styles and periods, activities) which are subdivided into thirty-three sub-facets or hierarchies. It is currently supported by the Getty Information Institute (see their Web page at http://www.gii.getty.edu/vocabulary/aat.html).

The *Thesaurus for Graphic Materials* is in two parts: *TGMI: Subject Terms* and *TGMII: Genre and Physical Characteristic Terms*. *TGMI* is less structured than *AAT*, lacking its faceted and highly hierarchical arrangement, though it does follow standard thesaural guidelines. *TGMI* provides a broader, though smaller, vocabulary than *AAT*, suitable for a general subject description of images. A detailed comparison of these two vocabularies is provided by Greenberg (1993).

Other controlled vocabularies have been developed for specific collections but, for many collections of images, particularly those on the Web, natural language indexing is preferred. Natural language may be in the form of text in which the image is embedded (newsphotos in newspapers, for instance), descriptions or captions accompanying it, or hypermedia links. For instance, Guglielmo and Rowe (1996) used natural language requests to query a database of historical images of aircraft and weapon

projects captioned with natural language text. By parsing and matching queries and captions, they were able to use natural language processing and inferencing techniques to answer queries such as "training missiles on a skyhawk."

In some contexts it seems logical to use images as surrogates for text in retrieval. A project at NASA's Johnson Space Center used a visual thesaurus to provide access to images from the Manned Space Flight Program, using images corresponding to those in a subject-oriented linguistic thesaurus (Seloff, 1990). This and other examples of visual thesauri are discussed by Hogan et al. (1991), who extend the concept of the visual thesaurus to the hypermedia environment, supporting browsing and searching through direct image links. This type of access corresponds to what Enser (1995) refers to as image retrieval in the VV mode—visual query, visual search.

Content-Based Indexing of Images

Content-based information retrieval (CBIR) refers to retrieval based on computer analysis of image content at the pixel level, automatically extracting such features as color, texture, and shape, locally or globally, from digital images. The CBIR systems currently available provide powerful retrieval engines for certain classes of query, although the developers have sometimes oversold their abilities, arguing that, since human indexing of image subject is prohibitively expensive, they propose to replace it by automatic indexing by color and texture. These systems are useful in some situations and no doubt will become more useful as their powers of interpretation become more sophisticated. The query categories proposed for them by Gudivada and Raghavan (1995) are color, texture, sketch, shape, volume, spatial constraints, browsing, objective attributes, subjective attributes, motion, text, and domain categories. Perhaps the capabilities that are currently best developed are retrieval by color, texture, and overall image similarity. Shape retrieval is most effective where solid images (clip art, trademarks) are queried. Domains where automatic indexing and retrieval have proven effective include face retrieval and fingerprints.

A realistic assessment of the state of the art of what they name visual information retrieval (VIR) is given by Gupta and Jain (1997). They indicate that systems providing information extraction from images still require some human image interpretation. The relative merits of concept- and content-based indexing are weighed by Flickner et al. (1995):

> Perceptual organization—the process of grouping image features into meaningful objects and attaching semantic descriptions to scenes through model matching—is an unsolved problem in image understanding. Humans are much better than computers at extracting semantic descriptions from pictures. Computers, however, are better

> than humans at measuring properties and retaining these in long-term memory. (p. 23)

Probably the best known such system, the Query by Image Content (QBIC) system developed by IBM, is commercially available and widely used (Flickner et al., 1995) (http://www.qbic.almaden.ibm.com). Image features are automatically extracted and stored in a database. Because of the problems in automatically outlining objects, manual and user-assisted techniques are used to identify shapes, though automated methods are available in some domains. Queries, which may be color and texture, user-drawn outlines, or sample images, are posed by sketching, selecting from a color palette, or selecting an image from a retrieved set as a further query. The QBIC system is currently being tested by the Department of Art and Art History at the University of California at Davis (Holt & Hartwick, 1994; Holt et al., 1997). They report better success with color and texture searches than with shape for content-based retrieval, with text searches preferred when artist or image is already known. Other similar systems include Virage (http://www.virage.com) and VisualSEEK (http://www.ctr.columbia.edu/~jrsmith/VisualSEEK/).

One of the more interesting developments in image indexing is the integration of concept- and content-based approaches using the information in descriptive text or captions to assist in the interpretation of the image. For instance, work by Srihari (1995) examines retrieval from a database of captioned newspaper photographs. Captions place constraints on the photographs, which help in identifying their content and the location in the image of the objects or individuals; however this information can often be interpreted only in the context of world knowledge, since human viewers are expected to recognize, for instance, President Clinton, or differentiate between Mr. and Mrs. Smith without spatial information. For example, it is hard to imagine the need for a caption as specific as "President Clinton (left) dancing with Hillary Clinton (right) at the Inaugural Ball." Research on the combination of textual and image sources of evidence for retrieval holds some promise in overcoming some of the disadvantages of text or image-based retrieval alone.

Image Indexing Environments—Case Studies

An examination of image collections on the WWW shows that it is not uncommon for access to be limited to a simple browsing approach. However, there are many collections in which indexing was used to improve access, either through a concept- or content-based technique. A few case studies will serve to illustrate the range of solutions that have been applied.

The Promenade system (McLean et al., 1994) was designed to provide access to a series of botanical images published in *Curtis Botanical Magazine*, an eighteenth-century compilation of images and text describ-

ing botanical samples collected by captains on voyages of exploration. The original intent was to use the natural language text of the descriptions as the index information, but the early printing, irregular typefaces, and nonstandard abbreviations made the text error-prone for OCR or human transcription, and a vocabulary and procedures for human indexing were selected. Since no existing vocabulary seemed well-suited to the historical and botanical nature of the image and textual materials, one was tailored to the image collection. This was an expensive solution, and the project could have benefited from content-based indexing techniques, then in their infancy, since retrieval by color was significant, and the clearly delineated botanical images would have allowed some degree of shape matching.

Two image database projects using controlled vocabulary are ELISE and Déjà Vu. The ELISE Project, funded by the European Commission, provides retrieval of full color images over a network (Black & Eyre, 1995). Initially two image collections, one from the Victoria and Albert Museum and one from Tilburg University Library in the Netherlands, were made available using both full-text descriptions and controlled vocabulary with the *AAT* as the source. Déjà Vu is an interface created for information retrieval systems in which users can browse through subject terms to find items that meet their information needs. The browsing process is facilitated by a knowledge structure in which subject terms are grouped based on the commonsense knowledge of library users in order to provide an interconnected browsing space. For example, when a user enters a search statement, the Broader Terms (BT), Narrower Terms (NT), Related Terms (RT), Notes, some relevant knowledge, and retrieved items will be displayed. The authors used the Library of Congress *Thesaurus for Graphic Materials* in this project (Gordon & Domeshek, 1998).

One of the more interesting applications of content-based retrieval systems is to databases of trademarks. While shape-based retrieval can be problematic in fine arts images with complex patterns of light and dark which make it difficult to extract individual shapes, trademarks are generally high-contrast shapes with good definition. The STAR system for trademark archiving and registration (Wu et al., 1995) is intended to allow searches for conflicting trademarks when a request is made for registration of a new image. The system ranks retrieved trademarks in order of similarity to the query trademark.

There are a number of instances on the Web of databases that are searchable using the QBIC software. For example, the Fine Arts Museum of San Francisco offers a QBIC search of a portion of its database comprising 3,000 Japanese prints. The similarity measure may be based on color percentages, color layout, texture, or a search may be customized using a color palette indicating the percentages desired of up to five colors (see their Web site at http://www.thinker.org/imagebase/index-2.html). IBM

offers demonstration searches of stamps, trademarks, and stock photos at their site at http://www.qbic.almaden.ibm.com/. The University of California at Davis study discussed above (Holt & Hartwick, 1994; Holt et al., 1997) can also be explored on their Web site at http://libra.ucdavis.edu/qbic.html.

Reports of the evaluation of image access systems are relatively rare in the literature. An exception is an evaluation of the Micro Gallery, a visitor information system at the National Gallery in London by Beaulieu and Mellor (1995). The system allows museum visitors to search the gallery collection by artist, historical atlas, picture type, and general reference. A combination of data collection methods was used to examine the impact of the interface features on search behavior, including questionnaires before and after the use of the system and direct observation with a talk aloud protocol.

Conclusion

With the increased availability of images comes the problems inherent in indexing them for retrieval. In order to develop solutions to these problems, more information is needed on why users search for images and how they intend to use them as well as how they pose their queries. Two approaches to image indexing have been developed and studied—concept-based and content-based. Concept-based indexing has the advantage of providing a higher-level analysis of the image content but is expensive to implement and suffers from a lack of interindexer consistency due to the subjective nature of image interpretation. Content-based indexing is relatively inexpensive to implement but provides a relatively low level of interpretation of the image except in fairly narrow and applied domains. To date, very little is known about the usefulness of the access provided by content-based systems, and more work needs to be done on user needs and satisfaction with these systems. An examination of a number of image database systems shows the range of techniques that have been used to provide intellectual access to image collections.

References

Armitage, L. H., & Enser, P. G. B. (1997). Analysis of user need in image archives. *Journal of Information Science, 23*(4), 287-299.

Beaulieu, M., & Mellor, V. (1995). The Micro Gallery: An evaluation of the hypertext system in the National Gallery, London. *New Review of Hypermedia and Multimedia, 1,* 233-260.

Black, K., & Eyre, J. (1995). The ELISE Project (electronic objects). In M. Collier & K. Arnold (Eds.), *ELVIRA 2* (Proceedings of the Second Electronic Library and Visual Information Research Conference, May 1995, De Montfort University, Milton Keynes, England) (pp. 70-78). London, England: Aslib.

Enser, P. G. B. (1995). Pictorial information retrieval. *Journal of Documentation, 51*(2), 126-170.

Enser, P. G. B., & McGregor, C. G. (1993). *Analysis of visual information retrieval queries* (British Library R & D Report No. 6104). London, England: British Library Board.

Flickner, M.; Sawhney, H.; Niblack, W.; Ashley, J.; Huang, Q.; Dom, B.; Gorkani, M.; Hafner, J.; Lee, D.; Petkovic, D.; Steele, D.; & Yanker, P. (1995). Query by image and video content: The QBIC System. *Computer, 28*(9), 23-31.

Gordon, A. S., & Domeshek, E. A. (1998). Déjà Vu: A knowledge-rich interface for retrieval in digital libraries. In *IUI 98* (International Conference on Intelligent User Interfaces, January 6-9, 1998, San Francisco, CA) (pp. 127-134). New York: Association for Computing Machinery Press.

Greenberg, J. (1993). Intellectual control of visual archives: A comparison between the *Art and Architecture Thesaurus* and the *Library of Congress Thesaurus for Graphic Materials. Cataloging & Classification Quarterly, 16*(1), 85-117.

Gudivada, V. N., & Raghavan, V. V. (1995). Content-based image retrieval systems. *Computer, 28*(9), 18-22.

Guglielmo, E. J., & Rowe, N. C. (1996). Natural language retrieval of images based on descriptive captions. *ACM Transactions on Information Systems, 14*(3), 237-267.

Gupta, A., & Jain, R. (1997). Visual information retrieval. *Communications of the ACM, 40*(5), 70-79.

Hastings, S. K. (1995). Query categories in a study of intellectual access to digitized art images. In T. Kinney (Ed.), *ASIS '95* (Proceedings of the 58th annual meeting of the American Society for Information Science, October 9-12, 1995, Chicago, IL) (pp. 3-8). Medford, NJ: American Society for Information Science.

Hogan, M.; Jorgensen, C.; & Jorgensen, P. (1991). The visual thesaurus in a hypermedia environment: A preliminary exploration of conceptual issues and applications. In D. Bearman (Ed.), *Hypermedia & interactivity in museums* (Proceedings of an international conference, October 14-16, 1991, Sheraton Station Square, Pittsburgh, PA) (pp. 202-221). Pittsburgh, PA: Archives and Museum Informatics.

Holt, B., & Hartwick, L. (1994). "Quick, who painted fish?"': Searching a picture database with the QBIC project at UC Davis. *Information Services & Use, 14*(2), 79-90.

Holt, B.; Weiss, K.; Niblack, W.; Flickner, M.; & Petkovic, D. (1997). The QBIC Project in the Department of Art and Art History at UC Davis. In C. Schwartz & M. Rorvig (Eds.), *Digital collections, implications for users, funders, developers, and maintainers* (Proceedings of the 60th Annual Meeting of the American Society for Information Science, November 1-6, 1997, Washington, DC) (pp. 189-195). Medford, NJ: Information Today.

Jorgensen, C. (1995). Classifying images: Criteria for grouping as revealed in a sorting task. In R. P. Schwartz, C. Beghtol, E. K. Jakob, B. H. Kwasnik, & P. Smith (Eds.), *Proceedings of the 6th ASIS SIG/CR classification research workshop* (October 8, 1995, Chicago, IL) (pp. 65-78). Chicago, IL: ASIS.

Jorgensen, C. (1998). Attributes of images in describing tasks. *Information Processing & Management, 34*(2/3), 161-174.

Keister, L. H. (1994). User types and queries: Impact on image access systems. In R. Fidel, T. B. Hahn, E.M. Rasmussen, & P. J. Smith (Eds.), *Challenges in indexing electronic text and images* (pp. 7-22). Medford, NJ: Learned Information.

Krause, M. G. (1998). Intellectual problems of indexing picture collections. *Audiovisual Librarian, 14*(2), 73-81.

Lancaster, F. W. (1998). *Indexing and abstracting in theory and practice* (2d ed). Urbana-Champaign: University of Illinois, Graduate School of Library and Information Science.

Layne, S. S. (1994). Some issues in the indexing of images. *Journal of the American Society for Information Science, 45*(8), 583-588.

Library of Congress. (1995). *Thesaurus for graphic materials.* Washington, DC: Cataloging Distribution Service, Library of Congress.

Markey, K. (1984). Interindexer consistency tests: A literature review and report of a test of consistency in indexing visual materials. *Library and Information Science Research, 6*(2), 155-177.

McLean, S.; Rasmussen, E. M.; & Williams, J. G. (1994). *Promenade:* Networked query and retrieval of horticultural images. In D. I. Raitt & B. Jeapes (Eds.), *Online Information 94* (Eighteenth International Online Information proceedings, 6-8 December 1994, London) (pp. 457-468). Oxford, England: Learned Information.

Panofsky, E. (1939). *Studies in iconology: Humanistic themes in the art of the Renaissance.* New York: Oxford University Press.

Rasmussen, E. M. (1997). Indexing images. *Annual Review of Information Science and Technology, 32,* 169-196.

Roberts, H. (1988). "Do you have any pictures of...?": Subject access to works of art in visual collections and book reproductions. *Art Documentation, 7*(3), 87-90.

Seloff, G. A. (1990). Automated access to the NASA-JSC image archives. *Library Trends, 38*(4), 682-696.

Shatford, S. (1986). Analyzing the subject of a picture: A theoretical approach. *Cataloging & Classification Quarterly, 6*(3), 39-62.

Sherman, C. R. (1987). ICONCLASS: A historical perspective. *Visual Resources, 4,* 237-246.

Srihari, R. (1995). Automatic indexing and content-based retrieval of captioned images. *Computer, 28*(9), 49-56.

Svenonius, E. (1994). Access to nonbook materials: The limits of subject indexing for visual and aural languages. *Journal of the American Society for Information Science, 45*(8), 600-606.

Wu, J. K.; Narasimhalu, A. D.; Mehtre, B. M.; Lam, C. P.; & Gao, Y. J. (1995). CORE: A content-based retrieval engine for multimedia information systems. *Multimedia Systems, 3*(1), 25-41.

Image Retrieval as Linguistic and Nonlinguistic Visual Model Matching

P. Bryan Heidorn

Abstract

This article reviews research on how people use mental models of images in an information retrieval environment. An understanding of these cognitive processes can aid a researcher in designing new systems and help librarians select systems that best serve their patrons. There are traditionally two main approaches to image indexing: concept-based and content-based (Rasmussen, 1997). The concept-based approach is used in many production library systems, while the content-based approach is dominant in research and in some newer systems. In the past, content-based indexing supported the identification of "low-level" features in an image. These features frequently do not require verbal labels. In many cases, current computer technology can create these indexes. Concept-based indexing, on the other hand, is a primarily verbal and abstract identification of "high-level" concepts in an image. This type of indexing requires the recognition of meaning and is primarily performed by humans. Most production-level library systems rely on concept-based indexing using keywords. Manual keyword indexing is, however, expensive and introduces problems with consistency. Recent advances have made some content-based indexing practical. In addition, some researchers are working on machine vision and pattern recognition techniques that blur the line between concept-based and content-based indexing. It is now possible to produce computer systems that allow users to search simultaneously on aspects of both concept-based and content-based indexes. The intelli-

P. Bryan Heidorn, Graduate School of Library and Information Science, University of Illinois, 501 E. Daniel, Champaign, IL 61820
LIBRARY TRENDS, Vol. 48, No. 2, Fall 1999, pp. 303-325

gent application of this technology requires an understanding of the user's visual mental models of images and cognitive behavior.

INTRODUCTION

To better understand the relationship between concept-based and content-based indexing in a volume such as this, it is useful to refocus and re-evaluate image indexing. An understanding of these techniques may be unified by examining how each relates to "visual mental models." From this perspective, image retrieval system work is an endeavor to create a concordance between an abstract indexing model of visual information and a person's mental model of the same information. All visual information retrieval research, from the computational complexity of edge detectors to national standards for museum indexing of graphical material, is an attempt to bring the indexing model and the user's mental model into line. All index abstraction, nonlinguistic or linguistic, may be classified by their success in matching the user's abilities. Borgman (1986) emphasizes that retrieval systems should be designed around "natural" human thinking processes. Index facet effectiveness is more dependent on the facets' harmonization of the facets with human cognition than on whether it is linguistic (concept-based) or nonlinguistic (content-based).

In describing the content of images in the realm of art, Panofsky (1955) distinguishes between pre-iconography, iconography, and iconology. Pre-iconographic content refers to the nonsymbolic or factual subject matter of an image. It includes the generic actions, entities, and entity attributes in an image. As an example, a pre-iconographic index may indicate that an image contains a stone (attribute), bridge (entity), and a river (entity). Iconographic content identifies individual or specific entities or actions. In the example, the bridge might be identified as the "Palmer Bridge" and the "Hudson River." The iconologic index would include the symbolic meaning of an image. The image might be indexed as "peaceful" or symbolizing "simpler times." The indexing that is appropriate depends on the type of subject matter that the searchers will eventually have in mind when they are doing a search.

This type of subject classification can be used to explain the strengths and weaknesses of content-based and concept-based indexing. Computers frequently perform content-based indexing. Computers can cost-effectively identify image attributes such as color, texture, and layout. Historically, limitations in computer algorithms have limited computer indexing to just a fraction of the pre-iconographics content. This, however, is changing, and the challenge for researchers and developers is to expand the functionality of the systems. Within limited contexts, computer indexing has been able to move into iconographic subject matter. For example, by exploiting information in picture captions in newspapers, a system may identify individuals in an image (Srihari, 1995). Other sys-

tems can identify and index objects such as trees or horses using low-level features such as texture and symmetry (Forsyth et al., 1996). Linguistic content-based indexing has traditionally been performed by humans. While it is expensive and time consuming, it is possible to create indexes for all three types of content matter described by Panofsky. Hastings (1995) demonstrated that, in some retrieval situations, searchers use a combination of both visual and verbal features. With current technology, this means the use of both content-based and concept-based techniques.

This article will focus on pre-iconographic indexing since this is the main area where content-based and concept-based techniques overlap. Content-based techniques may be used effectively where the computer can extract and synthesize features, attributes, and entities in images that are consistent with human understanding of the images. The computer must model the image in a way that is isomorphic (but not identical) to the human model of the image. Human indexers and searchers must also shape representations or mental models of the images if the indexer is to produce a functional index. In order to demonstrate the importance and pervasiveness of this process, this article will explore two aspects of indexing: color and object naming (shape). The first section will discuss the cognitive and social processes that give rise to the visual mental models that are shared by indexers and searchers. The next section explains what is meant by mental models in this context. Following this is a discussion of the representation of objects and shapes in visual mental models and then how both content-based and concept-based indexes capture (or neglect) aspects of these models. This is followed by a discussion of color in mental models and then discussion of the approaches to concept-based and content-based indexing by color.

Image Access as a Sociocognitive Process

Imagine an image of a bridge at sunset on a winter day. What color is the sky? Is there a name for the color? What objects are in the image? Are they important? Is the sun visible or has it already descended below the horizon? If you wanted to store this image with 100,000 others, how would you find it again? How would you describe it so that someone else could find it? Would words be enough? The answer to all of these questions depends on personal history and cultural expectations.

The act of indexing and accessing images from a database is a sociocognitive process grounded in both biology and experience. The term "sociocognitive" here means a combination of the social aspects of cognition as well as the individual aspects of mental life. Cognition refers to all processes involved in the perception, transformation, storage, retrieval, manipulation, and use of information by people. Of particular interest here will be those aspects of cognition that are called mental models. In a social context, we often wish to communicate our thoughts to

others. We frequently do this with language but also through our postures, gestures, or hand drawn illustrations or, for the gifted, through works of art. Communication between people is an act of one person referencing and changing the representations used in the cognition of another person, what they are thinking about, and even how they are thinking. In this context, indexing is a form of communication between the indexer and the people who will search for images in a collection. The indexer must rely on both shared cognitive heritage and social conventions to represent salient aspects of an image in the indexing scheme. The searchers, in using the index, must express their interests in the same language that was used by the indexers.

In the first paragraph of this section of the article, you were asked, through natural language, to create a "visual mental model" or "image" in your mind. Each reader's image is different, but certainly there are aspects of the image that are shared among readers. Some of these aspects may be based on the shared biology of our vision systems (most of us can imagine color), and some shared aspects may be attributable to our shared experience. We all know what bridges are without having been born with that knowledge. Some aspects of the visual mental model are easily described with natural language or verbal tags. Other aspects seem to defy simple linguistic description. "Although grammars provide devices for conveying rough topological information such as connectivity, contact, and containment, and coarse metric contrasts such as near/far or flat/globular, they are of very little help in conveying precise Euclidean relations: a picture is worth a thousand words" (Pinker & Bloom, 1995, p. 715).

This linguistic versus nonlinguistic contrast parallels concept-based and content-based indexing techniques. Understanding these mental models of images and how we can communicate information about them can enlighten us regarding content-based and concept-based indexing. Shera (1965) identified prerequisites for constructing a framework for indexing (an indexing vocabulary). These include an understanding of language and the communication process as well as an understanding of the relationship between human thought and mechanisms for recording thoughts such as language (p. 56). Indexers and system designers need to understand human cognition and communication in order to produce good indexes. The shared cognitive abilities and shared experience serve as the basis for this communication. These shared attributes may also arise from general world experience as in the earlier sunset example. Other attributes may arise from specialized training such as when an architect uses the *Art and Architecture Thesaurus* (Barnett & Petersen, 1989) to access a cultural heritage image collection or when a botanist uses the language in an identification key to label a specimen. In both cases, these cognitive attributes are learned in a social context.

In this discussion, the term "sociocognitive" is intended in its broadest sense. The social context here includes the conventions that allow indexers and searchers to learn common terminology, the natural and synthetic ontologies for image description. It is these aspects of the social environment that exist in a deep interplay with the shared cognitive abilities, biases, and frailties of the image access community. Cognitive abilities include not only a "higher" cognitive process but also the perceptual experience that is often the object of the "higher" cognitive processes. In this article, we do not focus on the social processes that indexers participate in to create indexing standards, although this is certainly important. The focus here is on the social environment that gives rise to the indexer's thoughts about images.

Jacob and Shaw (1999) introduce a sociocognitive perspective on representation. From their perspective and the perspective of this article, language and communication influence the organization of knowledge at both the individual and social level. Social processes lead to the creation of a shared vocabulary to describe a field. However, the Jacob and Shaw treatment is primarily limited to linguistic constructs: "[R]epresentation is primarily linguistic, the development of truly effective systems of retrieval must include a thorough appreciation of how language is used in the social processes of communicating knowledge" (p. 131).

When describing images, however, content-based indexing techniques introduce nonlinguistic forms of indexing (and communication), so this sociocognitive perspective must be extended to include nonlinguistic processes (such as color and texture maps). For images, it is clear that descriptions are grounded first in the perceptual abilities of indexers and searchers.

This does not diminish the critical role of natural language in image description. The creation of a vocabulary to describe images is a Darwinian adaptation and is universal to the species. This language learning is a sociocognitive process. For example, the perception of color is physical, but the color names are arrived at through a social process. There are millions of colors that people can distinguish (Bruner, Goodnow, & Austin, 1956, p. 1) but only some are named. An information retrieval system designer must decide if a collection should be indexed using the unlabeled colors (i.e., color histograms) or using labeled category names such as "red," "green," or "blue." The designer may choose to use both nonlinguistic and linguistic approaches. The decision must be made on both sociocognitive and technical grounds. In the mental image of a bridge at sunset, it might be reasonable to apply the label "red" for the sky. However, the colors in an actual sunset, or in our mental images, may defy our language skills. Figure 1, "Sunset, Palmer Bridge, New York"[1] is a digital image from the American Memory Collection at the Library of Congress (Detroit Photographic Co., c1900).[2] In this image, the sky's color does

not have a name with which many people would agree. The designer must decide if the users have a word to describe the particular shade of a sunset that is needed to complement the color of a car in an automobile advertisement. Nonlinguistic, content-based color retrieval is provided in current commercial and research image database systems such as Virage (Gupta et al., 1997), VisualSEEK (Smith & Chang, 1996a), QBIC (Niblack et al., 1992; Flickner et al., 1995), and Photobook (Pentland, 1993). These include, among others, color swaths, color mixing interfaces, perceptually significant coefficients, and color similarity matching as discussed in the section on models of color.

Figure 1. "Sunset, Palmer Bridge, New York."

Mental Models of Images

When a person is searching for an image in a collection, they may be thought of as searching for images that match a mental model of the image being sought. The mental model of the target may change during the course of the retrieval session, but this does not influence the fact that there is a dynamic mental model or how the model is constructed. If the collection is small enough, the searcher may browse the images looking for one that matches the mental model. When the collection becomes too large for efficient browsing, other search strategies must be employed. In the realm of image databases, the searcher may use an index. The appropriate nature of the index is governed by the nature of the mental representation. All current indexing techniques, both manual and auto-

matic, linguistic and nonlinguistic, are attempts to make aspects of the mental representation explicit and match these aspects to the images in the collection. As depicted in Figure 2, aspects of the visual world are abstracts by the searcher and the indexer. The indexer must select aspects of the abstraction that are shared by the indexer and searcher and code them into the index so that the index itself is an abstraction of the visual world. Because of the nature of this matching process and the complexity of the visual mental models, neither concept-based nor content-based indexing alone is sufficient to support an effective retrieval system. The best aspects of these approaches to indexing need to be identified and integrated.

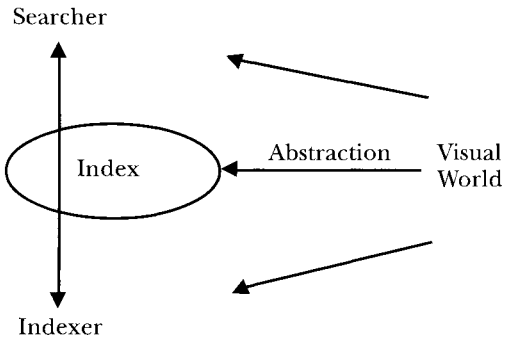


Figure 2. Index as Communication.

There are two types of correspondence that must exist between people and an image retrieval system—mental-model-to-index correspondence and cognitive-model-to-interface correspondence. The mental model-to-index correspondence is the degree to which a particular indexing facet is in harmony with the cognitive/perceptual models and predispositions of the searcher. The cognitive-model-to-interface correspondence is the degree of agreement between the searcher's cognitive/perceptual models and the ability to express these in the interface. This applies not only to the representation of the index in the interface but also to the user's expectations and mental models about how interfaces work (Borgman, 1986).

It is important, then, to consider the nature of the visual mental representations and their relationship to the physical world. Mental models of images represent, at least, perceptible aspects of the world that they represent (Johnson-Laird, 1983, p. 157). For the purposes of this analysis, it does not matter whether the representation is like an image in one's mind (Kosslyn, 1980; Paivio, 1971), is propositional (Pylyshyn, 1973; Palmer, 1975), or both. In either case, these models are abstractions of the visual

world and are not actual images since this would require the existence of a homunculus to observe these models.

These mental models possess some isomorphic relationship to the visual world. When people imagine a bridge at sunset, they are constructing an active mental model in working memory out of long-term memory traces. The processes involved in perception determine the contents of a long-term model. That is, the model of an image begins with its perception. The stages of processing from the outside world to long-term memory include sensory detection, pattern recognition, short-term memory, and long-term memory. In the visual system, sensory detection is the conversion of light into nerve impulses. Only light of very particular wavelengths can be detected but, as discussed later in this section, these impulses can serve as the basis for distinguishing millions of colors. Long-term mental models may contain representations of these colors, and people may wish to search image collections based on them. Content-based indexing methods for color representation support this spectrum-like aspect of the mental model.

The next stage of perception is pattern recognition. Our visual systems are trained from birth to recognize patterns in our environment. We have physical apparatus and training which allows us to detect edges, surfaces, depth, motion, and other aspects of the environment. This recognition is sometimes associated with the linguistic label for the pattern, but linguistic labels are not necessary. So we may recognize a particular pattern as being a cat and apply that label (bringing with it an association to a "cat" category in memory). We can also recognize objects for which we have no name. For example, in a zoo or in a forest, we may see an animal that we have never seen before. The fact that we have no name for it does not mean that we do not recognize it and remember it. In fact, this type of pattern recognition is the basis for a significant application of image databases. It is possible to identify animals, plants, or archeological objects by finding like objects in an image collection. Concept-based indexing techniques may be used where an object or pattern is named. Content-based techniques may be used where no name is available for at least some of the database searchers. Most thesauri for graphical materials, such as the *Art and Architecture Thesaurus (AAT)* (1994) and the Library of Congress' *Thesaurus for Graphic Materials* (1995), are examples of the concept-based approach. In these resources, all objects and patterns have labels.

The next stage in visual processing is short-term memory or working memory. The human memory system is frequently conceptualized as having two components: short-term memory and long-term memory. Two main properties differentiate the storage mechanisms. Short-term memory is limited in both size and duration. It is the mechanism used to remember information that may be forgotten immediately after use.

This might include a phone number or a URL. In some situations, short-term memory is better named "working memory." It includes the mechanisms that allow us to manipulate mental representations including mental images (as discussed in the next paragraph). Short-term memory is the procedure used to combine information from a visual scene with long-term memories. Long-term memory does not have either the duration or size limits of the short-term memory. Long-term memory is, however, very susceptible to distortion. One particular memory of an event can easily "mix" with prior memories and expectations. From the information processing perspective, memories in long-term memory must be moved to working memory before one is able to act on the memory. During an image retrieval task, the searcher will form a mental model of the target image in working memory. This model will be dynamic. Information from sensory input and from long-term memory can move into working memory. The sensory input might alter details of the model as near misses are encountered or as the user interface suggests options. Likewise, details of a scene may be filled in from long-term memory as the need arises.

People activate visual mental models or construct them from memory and then use them as a basis for comparison of images in a database. In some situations, these models behave as if they were three-dimensional representations very close to those used in perception. There are retrieval mechanisms that exploit the image-like qualities of images. These mechanisms allow the use of image qualities directly without the intervention of linguistic labels. These include color-wheels, color spaces, texture menus, sketching shapes by hand, by example-based searching, and other techniques. Indexing techniques sometimes treat images as if they were lists of attributes, but the mental models of the users are more like pictures in the mind. Sometimes the searcher can "read-off" individual attributes from those mental images, but the mental image itself is a more integrated whole.

There is psychological evidence for this integrated view of visual mental models. Like physical objects, these models take time to rotate mentally (Cooper & Shepard, 1973; Shepard, 1978). When subjects are asked to compare rotated versions of the same object to verify that they are the same, the time required to do the comparison is proportional to the angular difference between the images. The larger the angular difference, the more time that is required to make the judgment. People also seem to scan these mental images as if they were images that their eyes are scanning. Kosslyn, Ball, and Reiser (1978) asked people to memorize highly schematic maps and then asked them to answer questions about the location of objects on the map. When the question arose about the location of an object, the time to reply varied with the distance that the object was from the prior location that they had been asked about. If the prior object

had been further away on the map, it took longer to answer than when the prior object had been nearby.

These mental models in the mind of an indexer or searcher can be descriptively sparse or rich depending on the situation. Some components of the model may be easily described linguistically, but other aspects might best be described or communicated by example or by images. The next four sections of this article will take a closer look at the attributes of shape and color in both mental models and in image indexes. Content-based and concept-based indexing will be related to each of these mental model attributes.

Shape in Visual Mental Models

Psychologists have attempted to understand how perception gives rise to cognition and understanding. Many of these theories propose the existence of perceptual primitives. These are sometimes used in content-based approaches. In object recognition, these primitives may be generalized cones (Marr, 1982), simple geometric solids called "geons" (Biederman, 1987), or primitive features (Treisman & Gelade, 1980). When people are asked to describe objects, they often do so in terms of their parts (Tversky, 1989; Tversky & Hemenway, 1984). These primitive elements are combined in particular configurations to represent complex objects. In order to recognize and remember a bridge, the visual system breaks the scene into simple (more easily distinguished) parts. In early visual processing, a bridge may be broken into wedge-like supports and a rectangular prism for the deck. Once the shape has been recognized as a bridge, this fact may be added to long-term memory in abstracted form. The level of abstraction depends on the requirements of the task the person is working on. Information about the form of the bridge, including the shape of the arches, may be stored, or these details may be lost and only a strong trace of the general concept for "bridge" will be stored in long-term memory. For an image retrieval system, this means the indexer should index on just those properties that people have access to. These properties are dependent on both the shared memories of the users and the task parameters that accentuate particular aspects of these memories.

In the fields of computer vision and image retrieval, systems are often devised in layers. Primitive features are extracted from a scene and then combined into more complex features. David Marr (1982) articulated a model of visual primitives and generalized cones that served as a basis for much current research. Marr attempted to model human vision from the retinal image to object recognition. The lower level features are used for the construction of the 2.5-dimensional sketch. This sketch contains attributes that allow for later processing without the necessity to deal with lower level features such as light variations and discontinuities attributable to occlusion.

Shape in Indexes

When people view shapes, the shapes are recognized independent of their location (translation), their orientation (rotation), and their scale. Recognition is relatively resistant to noise. Variations in lighting and small occlusion do not interfere significantly. If a bug is missing a leg, it is nonetheless a bug. Some features selected by the visual apparatus are considered more important than others in defining similarity. Ideally, content-based algorithms that define shape similarity should behave in a manner consistent with human expectations and with the techniques that people use to define shape. The problem for content-based indexing is that current computational techniques do not have all of these properties. They tend to be effective at finding individual visual features, but the features frequently are not the same ones that people would recognize. They also tend to be poor at integrating the features to classify or recognize more complex objects. They are effective in recognizing straight lines and arches but not at recognizing that a particular combination of lines, edges, and colors is a bridge.

Still, this type of processing is the goal of many research systems. Consistent with the machine-vision tradition, content-based image retrieval systems model low-level feature-based information such as color, texture, and rough shape. These are used as evidence for the existence of higher level features or objects. Mehrotra (1997) provides a framework for understanding the levels of abstraction that may exist between an image and the viewer. The graphic of this model is reproduced in Figure 3. At the lowest level, there are *image features.* In that model, these features include color histograms, boundary segments, texture, and other "simple" features.

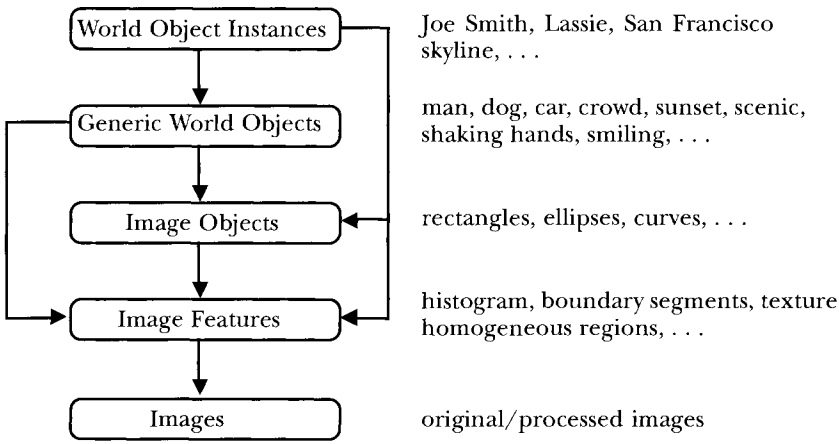


Figure 3. Levels of Abstraction in Museums (from Mehrotra, 1997).

Image objects, the next level of abstraction, are derived from collections of *image features*. Image objects include image regions, rectangles, and basic forms. The next level of abstraction is the *generic world object* such as man, dog, cat, or a smile. These include objects or categories to which many objects may belong. *World object instances* represent the next level of abstraction. These include objects for which there is one instance in the world that relates to the representation.

The concept-based approach to shape indexing focuses almost solely on generic world objects and world object instances. The indexer manually selects the relevant objects in an image and assigns keywords to the image. This linguistic tag approach is the primary means of image indexing in use today. The problems with this approach include expense, synonymy, and coverage. The manual operation requires a great deal of human effort to assign consistent tags and is therefore expensive. Another problem is that it is possible to describe an image many different ways. Even for textual material, it is difficult to select index terms that will be obvious to later searchers. This aspect of the problem is somewhat alleviated by the use of controlled vocabularies and thesauri, but then users are required to know that vocabulary. The problem is compounded in images. Sometimes users may wish to search for objects for which they have no name at all. This situation is not uncommon in that the image database is being used to facilitate object identification, as is the case with electronic field guides for the identification of plants and animals. Finally, the issue of coverage overlaps with that of expense. Indexers cannot normally create an entry for every object in an image. It is also very rare that an indexer has the time to index lower-level features such as the color or texture of an object or region. Consequently, when using the manual method of indexing, many objects and regions go unindexed.

These problems with content-based indexing can be demonstrated with the *AAT* by examining potential indexing options for the bridge in Figure 1 ("Sunset, Palmer Bridge, New York"). Part of the "IDNO 7836; TERM bridge" entry from the *AAT* is included in Figure 4. The index might be deepened by including the types of bridge that may apply but following the LINK entry "Bridge, stone" or adding entries for "IDNO 7838; TERM arch bridges" or "IDNO 7898; TERM single span bridges," if this is indeed a single span bridge. The parts of the object might be specified by following the related term of "RT <bridge elements>" from the "bridge" entry. Depending on the intended use of the index, the term "IDNO 994; TERM arch" could be included. The indexer would also need to decide which other objects in the image need to be indexed such as "IDNO 132410; TERM trees," "IDNO 8707; TERM river" (or perhaps "IDNO 8699; TERM stream" or "IDNO 11772; TERM water"), and "IDNO 133101; TERM winter." The correct index terms are not determined by the *AAT* but by the indexer's sociocognitive perspective on the intended

IDNO 7836
TERM bridges (built works)
ALT ALTERNATE bridge (built work)
BT <transportation structures by form>
RT <bridge elements>
SN SCOPE NOTE: Structures spanning and providing passage over waterways, topographic depressions, transportation routes, or similar circulation barriers
LINK bridges
LINK Bridges, aluminum
LINK Bridges, brick
LINK Bridges, concrete
LINK Bridges, iron and steel
LINK Bridges, masonry
LINK Bridges, plate-girder
LINK Bridges, prefabricated
LINK Bridges, stone
LINK Bridges, tubular
LINK Bridges, wooden

Figure 4. "Bridge" Entry in the *Art & Architecture Thesaurus.*

use. Even with this effort, these linguistic markers alone may be inadequate. Content-based techniques might facilitate some of the indexing and access.

Forsyth (1999) describes a system that represents a midpoint between content-based and concept-based approaches. This system uses a set of low-level image properties to infer the existence of objects. For example, an area of an image with a skin-like color, extended bilateral image symmetry, and nearly parallel sides might be a human limb. In a similar manner, it might be possible to build a bridge detector based on low level features. For example, the arched bridge in the image "Sunset, Palmer Bridge, New York" could be detected as a large dark area (stone) with a prominent arch(es) bounding the bottom and a near horizontal vertical line bounding the top. Based on the results of the bridge detector, "bridge" can be entered into the database along with a value representing the certainty of the classification. This technique borrows heavily from vision research and has the goal of being able to perform concept-based indexing at least within limited domains. The weakness of the technique is that detectors must be built for all objects of interest. The detector for arched bridges might not generalize to other bridge types, such as suspension bridges, requiring the construction of another detector. Detectors for rivers, trees, and sunsets would need to be constructed.[3]

In some situations, it may be possible to introduce a visual thesaurus. This type of thesaurus represents the choices visually rather than in natural

language as is the case with typical thesauri (Hogan et al., 1991). This allows people to "see" the visual indexing structure of a collection.

The techniques used in most content-based systems are aimed at a lower level in Mehrotra's hierarchy and stop with image features. The main techniques include template matching and edge abstraction matching. Most techniques of this type are two-dimensional projections of three-dimensional objects and suffer from perspective dependence. A query is constructed by using an example image from which the system may extract a shape outline or by hand sketching the desired shape. Rotation and scaling can also cause a mismatch. For example, the profile of a bridge looking from the road crossing is very different from the profile from the river the bridge crosses. Likewise, the same bridge from two different distances may produce different results. Current research is aimed at eliminating these types of limitations.

A general discussion of template matching can be found in Forsyth (in this issue of *Library Trends*). The System Query by Image Content (QBIC) (Barber et al., 1992; Niblack et al., 1992; Flicker et al., 1995) is a typical example of this approach. In template matching, a shape is normalized through translation, rotation, and scaling to produce an easily comparable standard form or template. These templates may be automatically extracted, but it is easier to have a user provide a sketch or outline of the desired object. The indexer, with the assistance of the computer, sketches the outline of objects of interest in an image. The system converts these outlines into templates by applying a standard rotation, scale, and translation. The system then stores the template as an index. In the sample image, the indexer would sketch the outline of the bridge. When users search the system, they may sketch the desired object. The system converts this sketch to a template and then compares this template to those in the index by counting the number of overlapping pixels. The greater the count, the higher the similarity. Allowing the indexer to add a name to the sketches could augment this technique. Unfortunately, the technique is sensitive to small variations in the images. In our sample image, the edges of the bridge are partly obstructed by trees. The indexer and searcher may choose different edge boundaries leading to a template mismatch. The same bridge from a different perspective would not be recognized or retrieved although scaling can be compensated for by QBIC. The advantage of the technique is that one algorithm applies to all objects. There is no need to create new detectors for each object of interest.

There are a number of edge abstraction techniques for classifying shape. These include, for example, turning angle descriptors, segmentation, and Fourier descriptors. Mehrotra and Gray (1995) describe a shape representation based on segmenting the edge of objects into straight-line segments. These segments are normalized for scale, rotation, and transla-

tion. Similarity is defined as the Euclidean distance between normalized points. The normalization helps to make the algorithm match human expectations, but the establishment of a start location and break points for the segmentation is problematic.

Another example of a boundary-based shape similarity approach is the Modified Fourier Descriptor (MFD) (Huang et al., 1997). This approach corrects faults in the Fourier Descriptor approach to produce a representation that is consistent in the face of transformations and noise.

All three of these approaches are weak in that they are not well-matched to human performance or expectations. They do not break objects into parts or other psychologically relevant features. Among these is the critical issue of dimensionality. Humans perceive two-dimensional images as three-dimensional. People combine the evidence in the image with long-term models in memory to produce three-dimension-like visual mental models (Hayward & Tarr, 1997). The QBIC template matching technique, the line-segment technique, and Fourier descriptors all act on two-dimensional projections of three-dimensional objects. Model-based computer vision research is focused on solving this projection problem.

A key issue is how any of these methods relate to users' mental models and how they operate at the interface level. If a user has a mental model and retrieval goal of a particular type of bridge at a particular orientation low-level feature, content-based techniques may be appropriate. If, however, these details are not relevant in the mental model or unspecified in the model then the concept-based approach is more appropriate. If the domain is narrow enough, this content indexing might be provided by automatic techniques such as those developed by Forsyth.

Color in Mental Models

Color is not light of a particular wavelength but rather it is combinations of light of different wavelengths. It is possible to produce the same perceived color through many combinations of wavelengths and intensity of light. The perception of color derives from the relative activation of three types of color receptors in the human retina. These receptors have highest sensitivity to wavelengths corresponding approximately to red, green, and blue. Red and green act as opponent colors, as do combinations of red and green receptors (yellow) and blue. The activation of one opponent color leads to the inhibition of the other opponent color—e.g., the perception of yellow stems from simultaneous moderate activation of both the red and green receptors. Brightness or intensity is encoded separately as a combination output of red and green receptors. Sharp and Philips (1997) provide a brief discussion of the neural aspects of vision. Hendee (1997) provides a discussion of the cognitive interpretation of color.

In a retrieval environment, multiple levels of abstraction may exist in both the index and in the minds of the searchers. Objects at all levels of abstraction sometimes have linguistic labels that are available to the searcher or indexer. For example, at the image feature level, a particular color may or may not have a color name associated with it. In a content-based index, a region's color might be stored as a color histogram with no linguistic label. The viewers may possess no words in their mental model to describe the color. There are 7 million discernible colors. Categorization and naming allows us to reduce this complexity. We cannot name them all (Bruner et al., 1956). The number of named primary colors may vary (in very systematic ways) from culture to culture as discussed below. Indeed, the meaning of the label may vary with the object to which it is applied. For example, the location on a color map of a *red apple* is different from the label for red in *red skin* (Clark, 1992, p. 370). Conceptual indexing would work best for the first and content-based techniques for the latter.

Some aspects of a model may be easily nameable and others may be difficult to assign labels to. The indexer must select the technique that is appropriate for each type of image element. Color names follow cross-cultural patterns. The indexing must follow these patterns or run the risk of producing confusing results to a user's queries. Lakoff (1987, pp. 24-40) discusses research into the relationship between color categories and color names. It is possible to assume that the assignment of color names to the spectrum is arbitrary. Different cultures might focus on different colors that are relevant in their environment and assign them names. Contrary to the arbitrary color hypothesis, Berlin and Kay (1969) demonstrated that there is a set of basic colors shared across cultures. These colors tend to have shorter names, are used more frequently, and there are about eleven of them. These are *black, white, red, yellow, blue, green, brown, purple, pink, orange* and *gray,* in that order. Of course the actual word used to represent the color differs but the colors are the same. When a language has only two basic color names, they are *black* and *white.* The other basic colors are grouped under these terms as *dark* or *bright* colors. When a language has a third basic color named it is *red.* When there is a fourth basic color term in a language it is usually *yellow, blue,* or *green.* Any one of these may be added first. Languages with six color terms will have the equivalents of *black, white, red, yellow, blue,* and *green.* The seventh color is *brown. Purple, pink, orange,* and *gray* are added next in no particular order. The obvious question is "Why" and is beyond the scope of this article. Interested readers, however, might start with Kay and McDaniel (1978) or Lakoff (1987).

The significance of this research for indexing is that if one is going to use linguistic labels for color names, these are the ones to use first. As is discussed in the next section, these labels are indeed included in the *AAT.*

There is additional structure to the human perception of color. Rosch (1973) studied focal and nonfocal color patterns. The first observation is that there are best examples of colors that cross cultures. So the *red* in one culture is the same *red* in other cultures. This is even true in cultures that have no basic color name for red. These colors are easier to learn. The focal colors act as cognitive reference points (Rosch, 1975). While these colors may have a privileged status in mental models, experience can play an important role in identifying the meaning of a term like "red." The meaning of red varies with context, so the meaning of red is different for a red apple, red skin of a sunburn, or a red sunset. Focal red is located at the center of a neighborhood of meaning for the color red (Clark, 1992, p. 371).

Color in Indexes

Using the concept-based approach, it is natural to map color names to particular objects within an image. There are a number of available controlled vocabularies available for this purpose. The *ATT* has a hierarchical color naming system. "IDNO: 131648 TERM: chromatic colors" acts as a base term, these being *pink, red, orange, brown, yellow, olive, yellow green, green, blue,* and *purple.* This does not quite match the basic color names described earlier but it is close. These *AAT* color terms serve as base terms for less prototypical colors. For example, *blue* is the base term for "IDNO: 129787 TERM: <intermediate blues>" and in turn "<intermediate blues>" is the base term for "IDNO: 130602 TERM: violet" as well as other related colors.

Colors are further categorized using the "use for" (UF) fields. The chromatic color thesaurus entries (*pink, red, orange,* and so on) do not contain UF fields except *yellow, green* that has a UF of *green,* and *yellow.* "IDNO: 129645 TERM: pale blue" is a type of *blue* and the terms such as *dull greenish blue* are mapped to *pale blue* through the UF field. So the sky in the bridge at sunset picture might be coded as 129645. Individuals will vary on their definition of these more obscure color names (but not the focal colors).

There are also achromatic colors defined in *AAT.* These are colors without hue and include black, white, and grays. This is consistent with the color space discussion of content-based indexing later in this section.

There are thousands of color names defined in the *AAT.* It takes a good deal of effort on the part of both indexers and searchers to arrive at the same term for the color of a particular sky. There is some support for color similarity through the thesaurus. If a user enters a particular color term, the system should be able to search for images or objects in images with this label. Should the search fail, the system should be able to automatically relax the matching constraints by moving up to the base term (e.g., *pale blue* to *blue*) and perform a search with that high level

term. If that fails, the system might use OR to group all terms having the base term of *blue.* This is much to ask of a system, but similar systems have been built. There exist other color name standards, such as the *National Bureau of Standards Dictionary of Color Names,* which provides thousands of color names and the *National Bureau of Standards/NBIC Color System,* that maps all colors into a small set of over 200 names.

If a retrieval environment were to require the use of more than the primary colors, it would be unreasonable to expect either indexers or searchers to have names for them. It is even more unlikely that they would agree on the names. In these situations, it might be more reasonable to allow users to directly specify color. An analytic approach might allow users to select different levels of red, green, and blue from window slider bars for example. The combination of the colors would reproduce all other colors. There are multiple problems associated with deciding on a color indexing system—e.g., no one solution can fit all needs. Frequently, a combination of approaches is needed, drawing from both what has been defined as content-based and concept-based techniques. A few of the central issues include the decision of what to index, the determination of color averages and color naming (both qualitative and quantitative). For the content-based approach, the color histogram is the favored method for representing average color. Color histograms are discussed elsewhere by Forsyth in this issue of *Library Trends.* These may be used to represent the overall image color, the color of regions, or the color of objects in increasing level of difficulty.

Digital images are composed of a series of points. The color of a particular point may be represented in either qualitative or quantitative terms. Indexing on a pixel level is not very useful in most cases. As indicated in the Mehrotra model in Figure 3 above, these individual elements may be collected together into homogeneous regions as *image features.* Unlike concept-based indexing, these regions need not correspond to individual objects or even object parts. These regions or "blobs" may be indexed independently. Smith and Chang (1996a, 1996b) discuss one approach to this problem. Regions of similar color are labeled with their location and color.

There are many approaches to color, but the Smith and Chang approach is useful for the purposes of explanation since it demonstrates some basic ideas and is computationally tractable. In this approach, color is represented in the HVS color space: hue, value, and saturation. There are many other color spaces, and these will be discussed later. Hue is the tint of what is typically pure color. Saturation is the amount of color mixing where fully mixed red, green, and blue appears as white. Value is the lightness or intensity. Colors are quantized into a small number (166) of color regions: eighteen hues, three saturations, three values, plus four grays. Colors that fall anywhere in a region are considered the same for

indexing purposes and for identifying regions. This quantization of color is reminiscent of the categorization of color performed by humans. The choice of eighteen hues is interesting in that it does not correspond to the eleven basic colors identified by Berlin and Kay (1969) and Rosch (1974). Eleven hues might match human expectations better without adding computational complexity to the approach.

So, in the content-based approach, regions of like color or "blobs" are indexed. At first glance, such a nonobject-oriented approach would not seem to correspond to the human experience of the same image. Humans, after all, recognize physical objects in images as in Mehrotra's "generic world objects" (man, dog, car, and so on). In practice, however, the technique is sometimes useful because, happily, the "blobs" do correspond to objects. In an image database of nature photography, yellow blobs in the middle of the frame frequently correspond to yellow flowers and yellow blobs in a collection of bird photos often correspond to yellow birds. There is, of course, a high error rate that increases with the heterogeneity of the image collection. The color quantization approach is useful for finding color regions, but an additional mechanism is needed to handle color similarity. In the concept-based or keyword approach to color matching, either the color of an image matches the color of the query or it does not. That is, if a sky is indexed with the keyword "blue," only the word "blue" in a query will match it. This does not match with human modeling of color. We know from Rosch's work on color prototypes that colors are not created equal, and that some colors may be better or worse members of the "blues" than others. "Blue" is closer to "light blue" and "bluish-green" than it is to "red." Content-based color similarity methods can be built which much more closely match these intuitions. Again using Smith and Chang's quantized color region approach as an example, the distance between two colors, or their similarity, can be defined as the number of steps that need to be taken in the quantized space to move from one color region to another. Hue is broken into eighteen regions. The first region might correspond to something like reds, the second region to oranges, and so on up to the last visible violet. Regions that are close to one another are close in color. The "orange" bin is distance one from "red," and the "violet" region is distance seventeen from "red." The same applies along the (3) saturation and (3) value axes. Color distance is the sum of the hue, saturation, and value distances. A user may search for a blue sky and have a relatively strong color match for the sky in the "Sunset, Palmer Bridge, New York" image.

There are other color spaces and similarity measures that more closely match human perception. Human color perception, however, has certain limits. Some wavelengths are simply not visible. This may be because the wavelength of light is beyond the range of our color receptors (retinal cones). Likewise, the intensity may be too low or too high. The designer

might choose to use a model that may represent all visible colors. The Commission Internationale de L'Eclairage Color Space (CIE) is such a representation. This is a three-dimensional color model that represents saturated colors (red, green, and blue) on outside edges of a bounded plane. Unsaturated colors are in the central area with white in the middle. Intensity is expressed on an axis orthogonal to the color plane. One boundary is black and the other full intensity. While this model represents all visible colors, it does not compensate for human processing of the RGB channels. The CIE-LAB model does bit mapping of the color space into complementary colors. There is a red-green axis, a yellow-blue axis, and a black-white axis as there is in the central processing of color in humans. *Munsell*, a U.S. standard, is another popular standard. The problem with these spaces is that it is sometimes difficult to map the standard RGB encoding used in monitors and scanners.

These color spaces have been constructed to capture important aspects of the human perception of color. Human and computer indexers may use them as a tool to describe aspects of an image. This use of the color spaces will be successful inasmuch as they are consistent with expectations and mental models of the users of the index.

Conclusion

Humans have evolved mechanisms that allow them to represent important aspects of the visual world. These visual mental representations are used on a daily basis to recognize objects and navigate through the world. Many aspects of these visual models predate the evolution of language. Language evolved to facilitate our ability to communicate with one another—i.e., facts about the world and our understanding of the world. Language has access to particular aspects of our visual mental models, allowing people to describe their interpretation of the world. In order for others to understand these descriptions, there must be a shared experience of the world and a shared vocabulary. The nature of both the visual mental models and the linguistic mechanism have a profound effect on how image retrieval systems should be built. Indexers may use language and this shared knowledge to create language-based descriptions of images in a collection. Computer algorithms are being developed that allow some parts of this linguistic indexing to be performed cost effectively by computers at least in narrow subject domains (Forsyth et al., 1996; Forsyth, 1999; Srihari, 1995, 1997). These computer systems are breaking down some of the distinctions that have existed between content-based and concept-based indexing.

Some aspects of the visual mental models are not easily described with natural language. As discussed in the section on color indexing, there are millions of human-discernible colors but relatively few color names. In some cases, content-based computational techniques can be

used to communicate information about these nonlinguistic aspects of the visual models. These techniques are used in systems such as Virage (Gupta et al., 1997), QBIC (Niblack et al., 1992; Flickner et al., 1995), VisualSEEK (Smith & Chang, 1996a), and Photobook (Pentland, 1993). Some systems, such as Photobook, attempt to select image properties that are particularly perceptually salient. Some of the mechanisms involved in the representation of shape and color are discussed in this article. No one content-based representational technique is likely to capture all of the important aspects of an image. The mental model of images has multiple aspects. The image features of different types are reflected in the different aspects of the mental models. Content-based and concept-based approaches to indexing are each better suited to different aspects of the models. Indexers may choose to use content-based or concept-based linguistic or nonlinguistic indexing depending on the demands of the tasks that will be performed by the users and what aspects of the visual mental models will be available to them.

NOTES

[1] Reproduced with permission from the Library of Congress, Prints and Photographs Division, Detroit Publishing Company Collection.

[2] It is useful to be able to refer to the color version of this image in the American Memory Collection. The image may be accessed through the Web by searching for the title at http://memory.loc.gov/ammem/detroit/dethome.html.

[3] Interestingly, detectors for trees and sunsets have been constructed (see Forsyth in this issue of *Library Trends*).

REFERENCES

Art and Architecture Thesaurus. (1994). Getty Art History Information Program, 2d. ed. New York: Oxford University Press.

Barber, R.; Cody, W.; Equitz, W.; Flickner, M.; Glasman, E.; Niblack, W.; & Petkovic, D. (1992). *Query by image content (QBIC) status as of 8/92.* Unpublished Technical Report RJ 89 (80237) September 14, 1992. IBM Research Division, Almaden Research Center, 650 Harry Road, San Jose, CA 95120-6099.

Barnett, P. J., & Petersen, T. (1989). Subject analysis and AAT/MARC implementation. *Art Documentation, 8,* 171-182.

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution.* Berkeley: University of California Press.

Biederman, I. (1987). Recognition by components: A theory of human image understanding. *Psychological Review, 94,* 115-147.

Borgman, C. L. (1986). The user's mental model of an information retrieval system: An experiment on a prototype online catalog. *International Journal of Man-Machine Studies, 24,* 47-64.

Bruner, J.; Goodnow, J.; & Austin, G. (1956). *A study of thinking.* New York: Wiley & Sons, Inc.

Clark, H. H. (1992). *Arenas of language use.* Chicago: University of Chicago Press.

Cooper, L. A., & Shepard, R. N. (1973). Chronometric studies of the rotation of mental images. In W. G. Chase (Ed.), *Visual information processing.* Orlando, FL: Academic Press.

Detroit Photographic Co. *"Sunset, Palmer Bridge, New York." c1900* (Touring Turn-of-the-Century America: Photographs from the Detroit Publishing Company, 1880-1920). Retrieved November 10, 1999 from the World Wide Web: http://memory.loc.gov/ammem/detroit/dethome.html.

Flickner, M.; Sawhney, H.; Niblack, W.; Ashley, J.; Huang, Q.; Dom, B.; Gorkani, M.; Hafner, J.; Lee, D.; Petkovic, D.; Steele, D.; & Yanker, P. (1995). Query by image and video content: The QBIC system. *Computer, 28*(9), 23-30.

Forsyth, D.; Malik, J.; Leung, T.; Bregler, C.; Carson, C.; Greenspan, H.; Fleck, M. (1996). Finding pictures of objects in large collections. In P. B. Heidorn & B. Sandore (Eds.), *Digital image access & retrieval* (Proceedings of the 33rd Annual Clinic on Library Applications of Data Processing, March 24-26, 1996, University of Illinois at Urbana-Champaign). Urbana-Champaign: University of Illinois, Graduate School of Library and Information Science.

Gupta A., & Jain R. (1997). Visual information retrieval. *Communications of the ACM, 40*(5), 70-79.

Hastings, S. K. (1995). Query categories in a study of intellectual access to digitized art images. In T. Kinney (Ed.), *ASIS '95* (Proceedings of the 58th annual meeting of the American Society for Information Science, October 9-12, 1995, Chicago, IL) (pp. 3-8). Medford, NJ: American Society for Information Science.

Hayward, W. G., & Tarr, M. J. (1997). Testing conditions for viewpoint invariance in object recognition. *Journal of Experimental Psychology-Human Perception and Performance, 23*(5), 1511-1521.

Hendee, W. (1997). Cognitive interpretation of visual signals. In W. R. Hendee & P. N. T. Wells (Eds.), *Perception of visual information* (pp. 149-175). New York: Springer-Verlag.

Hogan, M.; Jorgensen, C.; & Jorgensen, P. (1991). The visual thesaurus in a hypermedia environment: A preliminary exploration of conceptual issues and applications. In D. Bearman (Ed.), *Hypermedia & interactivity in museums* (Proceedings of an International Conference, October 14-16, 1991, Sheraton Station Square, Pittsburgh, PA). Pittsburgh, PA: Archives & Museum Informatics.

Huang, T. S.; Mehrotra, S.; & Ramchandran, K. (1997). Multimedia analysis and retrieval system (MARS) project. In P. B. Heidorn & B. Sandore (Eds.), *Digital image access and retrieval* (Proceedings of the 33rd Annual Clinic on Library Applications of Data Processing, March 24-26, 1996, University of Illinois at Urbana-Champaign). Urbana-Champaign: University of Illinois, Graduate School of Library and Information Science.

Jacob, E., & Shaw, D. (1999). Sociocognitive perspective in representation. *Annual Review of Information Science and Technology, 33*, 3-57

Johnson-Laird, P. N. (1983). *Mental models.* Cambridge, MA: Harvard University Press.

Kay, P., & McDaniel, C. (1978). The linguistic significance of the meanings of basic color terms. *Language, 54*(3), 610-646.

Kosslyn, S. M. (1980). *Images and mind.* Cambridge, MA: Harvard University Press.

Kosslyn, S. M.; Ball, T. M.; & Reiser, B. J. (1978). Visual images preserve metric spatial information: Evidence from studies of visual scanning. *Journal of Experimental Psychology: Human Perception and Performance, 4*, 47-60.

Lakoff, G. (1987). *Women, fire, and dangerous things.* Chicago: University of Chicago Press.

Library of Congress. (1995). *Thesaurus of graphic materials* (comp. & ed. the Prints & Photography Division, Library of Congress). Washington, DC: Library of Congress Catalog Distribution Service.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* San Francisco: W. H. Freeman.

Mehrotra, R. (1997). Content-based image modeling and retrieval. In P. B. Heidorn & B. Sandore (Eds.), *Digital image access and retrieval* (Proceedings of the 33rd Annual Clinic on Library Applications of Data Processing held March 24-26, 1996, at the University of Illinois at Urbana-Champaign). Urbana-Champaign: University of Illinois, Graduate School of Library and Information Science.

Mehrotra, R., & Gray, J. E. (1995). Similar-shape retrieval in shape data management. *IEEE Computer, 28*(9), 57-62.

Niblack, W.; Barber, R.; Equitz, W.; Flickner, M.; Glassman, E.; Petkovic, D.; Yanker, P.; Faloutsos, C.; & Taubin, G. (1992). The QBIC project: Querying images by content using color, texture, and shape. In A. A. Jamberdino & W. Niblack (Eds.), *IMAGE storage and retrieval systems* (Proceedings of the SPIE—the International Society for Optical Engineering) (vol. 1662, pp. 173-181). Bellingham, WA: SPIE.

Paivio, A. (1971). *Imagery and verbal processes.* New York: Holt, Rinehart and Winston.
Palmer, S. E. (1975). Visual perception and world knowledge: Notes on a model of sensory-cognitive interaction. In D. A. Norman, D. E. Rumelhart, & LNR Research Group (Eds.), *Explorations in cognition.* San Francisco: Freeman Press.
Panofsky, E. (1955). *Meaning in the visual arts.* Garden City, NY: Doubleday Anchor Books.
Pentland, A.; Picard, R. W.; & Sclaroff, S. (1993). Photobook: Content-based manipulation of image databases. *International Journal of Computer Vision, 18*(3), 233-254.
Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences, 13*(4), 707-784.
Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin, 80,* 1-24.
Rasmussen, E. M. (1997). Indexing images. *Annual Review of Information Science and Technology, 32,* 167-196.
Rosch, E. (1973). Natural categories. *Cognitive Psychology, 4,* 328-350.
Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology, 7,* 532-547.
Sharp, P., & Philips, R. (1997). Physiological optics. In W. R. Hendee & P. N. T. Wells (Eds.), *Perception of visual information* (pp. 1-32). New York: Springer-Verlag.
Shepard, R. N. (1978). The mental image. *American Psychologist, 33,* 125-137.
Shera, J. H. (1965). *Libraries and the organization of knowledge.* Hamden, CT: Archon Books.
Smith, J. R., & Chang, S-F (1996a). VisualSEEK: A fully automated content-based image query system. In *Proceedings of the ACM International Conference on Multimedia* (November 1996). Boston: Association for Computing Machinery.
Smith, J. R., & Chang, S-F. (1996b). Tools and techniques for color image retrieval. In *Proceedings Storage & Retrieval for Image and Video Databases IV* (Vol. 2670). San Jose, CA: IS&T/SPIE.
Srihari, R. (1995). Automatic indexing and content-based retrieval of captioned images. *IEEE Computer, 38*(9), 49-56.
Srihari, R. (1997). Using speech input for image interpretation, annotation, and retrieval. In P. B. Heidorn & B. Sandore (Eds.), *Digital image access and retrieval* (Proceedings of the 33rd Annual Clinic on Library Applications of Data Processing held March 24-26, 1996, at the University of Illinois at Urbana-Champaign). Urbana-Champaign: University of Illinois, Graduate School of Library and Information Science.
Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12,* 97-136.
Tversky, B. (1989). Parts, partonomies, and taxonomies. *Developmental Psychology, 25*(6), 983-995.
Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General, 113*(2), 169-191.

Computer Vision Tools for Finding Images and Video Sequences

D. A. FORSYTH

ABSTRACT

VERY LARGE COLLECTIONS OF IMAGES are now common. Indexing and searching such collections using indexing languages is difficult. Computer vision offers a variety of techniques for searching for pictures in large collections. Appearance methods compare images based on the overall content of the image using such criteria as similarity of color histograms, texture histograms, spatial layout, and filtered representations. Finding methods concentrate on matching subparts of images, defined in a variety of ways, in the hope of finding particular objects. These ideas are illustrated with a variety of examples from the current literature.

INTRODUCTION

Some of the image collections described in Enser (1995) contain tens of millions of items. Accessing a collection of pictures is difficult because it is hard to describe a picture accurately. Indexing a large collection by hand involves a substantial volume of work. Furthermore, there is the prospect of having to re-index sections of the collection—e.g., if a news event makes a previously unknown person famous, it would be beneficial to know if the collection contained pictures of that person. Finally, it is very difficult to know what a picture is about—e.g., "the engineer requests an image of a misaligned mounting bracket . . which only exists as an image described by a cataloguer as astronaut training . . . " (Seloff, 1990, p. 686). These observations indicate that it would be ideal to have automated tools that can describe pictures and find them based on a description. IBM's seminal QBIC (Query by Image Content) system demonstrated

D. A. Forsyth, Computer Science Division, University of California at Berkeley, Berkeley, CA 94720
LIBRARY TRENDS, Vol. 48, No. 2, Fall 1999, pp. 326-355

that such tools could be built (a comprehensive description of QBIC appears in the system). This review will indicate the different approaches that computer vision scientists have taken in building tools—rather than listing available tools—and indicate the kind of tasks these tools will facilitate. Some ideas from computer vision will be introduced, but readers who would like a deeper discussion of the topic should consult texts by Forsyth and Ponce (in press) and Trucco and Verri (1998).

What Do Users Want?

The most comprehensive study of the behavior of users of image collections is Enser's work on the Hulton-Deutsch collection (Armitage & Enser, 1997; Enser, 1993, 1995). This is a collection of prints, negatives, slides, and the like used mainly by media professionals. Enser (1993) studied the request forms on which client requests are logged; he classified requests into four semantic categories, depending on whether a unique instance of an object class is required or not and whether that instance is refined. Significant points are that the specialized indexing language used gives only a "blunt pointer to regions of the Hulton collections" (p. 35) and the broad and abstract semantics used to describe images. For example, users requested images of hangovers, physicists, and the smoking of kippers. All these concepts are well beyond the reach of current image analysis techniques. As a result, there are few cases where one can obtain a tool that directly addresses a need. For the foreseeable future, the main constraint on the design of tools for finding images will be our limited understanding of vision.

The Hulton-Deutsch collection is used largely by book, magazine, and newspaper publishers; Enser suggests that the most reliable measure of the success of their indexing system is that the organization is profitable. This author is unaware of any specific areas of computer vision technology that are profitable, though Virage appears to be thriving (the company is described at: http://www.virage.com. A description of their technology appears in Hampapur et al., 1997). The main source of value in any large collection is being able to find items. Potential application areas include:

- *military intelligence:* vast quantities of satellite imagery of the globe exist, and typical queries involve finding militarily interesting changes—e.g., concentrations of force occurring at particular places (e.g., Mundy, 1995, 1997; Mundy & Vrobel, 1994).
- *planning and government:* satellite imagery can be used to measure development, changes in vegetation, regrowth after fires, and so on (see, for example, Smith, 1996).
- *stock photo and stock footage:* commercial libraries—which often have extremely large and very diverse collections—sell the right to use particular images (e.g., Armitage & Enser, 1997; Enser, 1993, 1995).

- *access to museums:* museums are increasingly releasing collections, typically at restricted resolutions, to entice viewers into visiting the museum (e.g., Holt & Hartwick, 1994a, 1994b; Psarrou et al., 1997).
- *trademark enforcement:* as electronic commerce goes, so does the opportunity for automatic searches to find violations of trademark (e.g., Eakins et al., 1998; Jain & Vailaya, 1998).
- *indexing the Web:* indexing Web pages appears to be a profitable activity. Users may also wish to have tools that allow them to avoid offensive images or advertising. A number of tools have been built to support searches for images on the Web using techniques described later in this article (e.g., Cascia et al., 1998; Chang et al., 1997b; Smith & Chang, 1997).
- *medical information systems:* recovering medical images "similar" to a given query example might give more information on which to base a diagnosis or to conduct epidemiological studies (e.g., Congiu et al., 1995; Wong, 1998).

What Can Tools Do?

There are two threads discernible in current work on finding images: one can search for images based either on the appearance of the whole image or on object-level semantics.

The central technical problem in building appearance tools is defining a useful notion of image similarity; the section of this article entitled "Appearance" illustrates a variety of different strategies. Image collections are often highly correlated so that it is useful to combine appearance tools with browsing tools to help a user browse a collection in a productive way—i.e., by trying to place images that are "similar" near to one another and offering the user some form of dialogue that makes it possible to move through the collection in different "directions." Building useful browsing tools also requires an effective notion of image similarity. Constructing a good user interface for such systems is difficult. Desirable features include a clear and simple query specification process, and a clear presentation of the internal workings of the program so that failures can be resolved easily. Typically, users are expected to offer an example image or to fill in a form-based interface.

It is very difficult to cope with high-level semantic queries ("a picture of the Pope kissing a baby") using appearance or browsing tools. Finding tools use elements of the currently limited understanding of object recognition to help a user query for images based on this kind of semantics at a variety of levels. It is not known how to build finding tools that can handle high-level semantic queries, nor how to build a user interface for a general finding tool; nonetheless, current technology can produce quite useful tools (see the later section entitled Finding).

APPEARANCE

Images are often highly stylized, particularly when the intent of the artist is to emphasize a particular object or a mood. This means that the overall layout of an image can be a guide to what it depicts so that useful query mechanisms can be built by searching for images that "look similar" to a sample image, a sketched sample, or textual specification of appearance. The success of such methods rests on the sense in which images look similar. A good notion of similarity is also important for efficient browsing, because a user interface that can tell how different images are, can lay out a display of images to suggest the overall structure of the section of the collection being displayed. I will concentrate on discussing appearance matching rather than browsing because of the similarity of the technical issues.

The simplest form of similarity—two pictures are similar if all their pixels have similar values—is extremely difficult to use because it requires users to know exactly how a picture is laid out. For example, if one were trying, with a sketch, to recover the Van Gogh painting of sunflowers on a table, it would be necessary to remember on which side of the table the vase of flowers lay. It is possible to build efficient systems that use this iconic matching (e.g., Jacobs et al., 1995), but the approach does not seem to have been adopted, probably because users generally are unable to remember enough about the picture they want to find. It is important to convey to the user the sense in which images look similar, because otherwise mildly annoying errors can become extremely puzzling.

Histograms and Correlograms

A popular measurement of similarity compares counts of the number of pixels in particular color categories. For example, a sunset scene and a pastoral scene would be very different by this measure because the sunset scene contains many red, orange, and yellow pixels and the pastoral scene will have a preponderance of green (grass), blue (sky), and perhaps white (cloud) pixels (see Figure 1). Furthermore, sunset scenes will tend to be similar; all will have many red, orange, and yellow pixels and few others. This color histogram matching has been extremely popular; it dates back at least to Swain and Ballard (1991) and has been used in a number of systems (Flickner et al., 1996; Holt & Hartwick, 1994b; Ogle & Stonebraker, 1995). The usefulness of color histograms is slightly surprising, given how much image information the representation discards; Chapelle, Haffner, and Vapnik (1999) show that images from the Corel collection[1] can be classified by their category in the collection using color histogram information alone.

There is no record in a color histogram of where colored pixels are with respect to one another. Thus, for example, pictures of the French and British flags are extremely similar according to a color histogram

measure—each has red, blue, and white pixels in about the same number; it is the spatial layout of the pixels that differs. One problem that can result is that pictures taken from slightly different viewing positions look substantially different by a color histogram measure (see Figure 2). This effect can be alleviated by considering the probability that a pixel of some color lies within a particular pixel of another color (which can be measured by counting the number of pixels at various distances). Requiring this measure to be similar provides another measure of image similarity. Computational details are important because the representation is large (Huang et al., 1997; Huang & Zabih, 1998). For small movements of the camera, these probabilities will be largely unchanged so that similarity between these color correlograms yields a measure of similarity between images.

Figure 1. Results from a query to the Calphotos collection that sought pastoral scenes, composed by searching for images that contain many green and light blue pixels. As the results suggest, such color histogram queries can be quite effective.

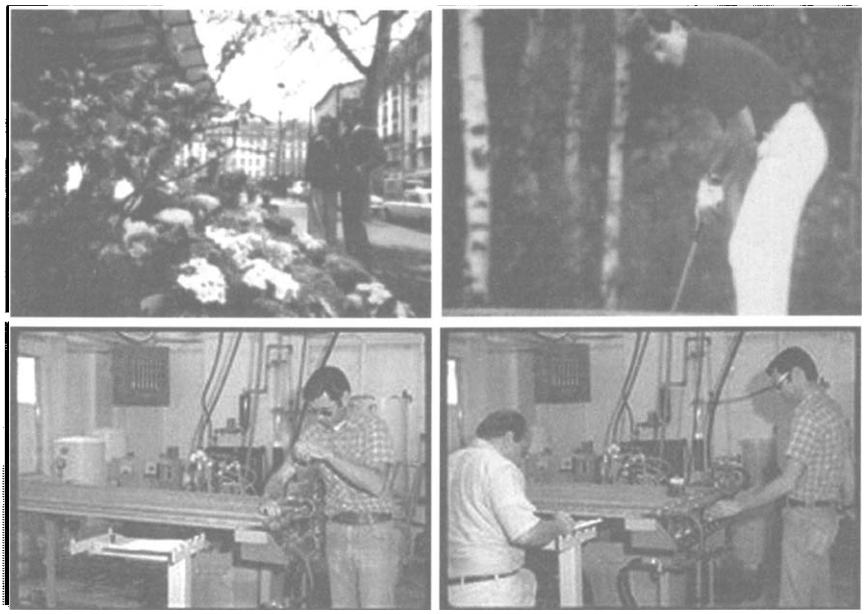


Figure 2. The top two figures have the same color histogram; the red patch on the golfer's shirt in the original print appears in the other image as brown looking flowers. These flowers were darker in color than they appear (their hue is changed somewhat by the fact that they are small and don't contrast strongly with their background), although not quite as dark in color as the shirt. However, in the scheme of color categories used, the colors are regarded as equivalent. The bottom two figures have similar content, but quite different color histograms; one person has a peach shirt and appears in only one figure and the one person wearing blue occupies more space in one than in the other. However, if one looks at a representation of the extent to which pixels of a given color lie near pixels of some other color, the pictures are quite similar—many blue pixels lie near either blue pixels, green ones, or white ones. This information is captured by the color correlogram. More information on color correlograms can be found in Huang and Zabih (1998). (Picture reproduced by kind permission of R. Zabih.)

Textures

Color histograms contain no information about the layout of color pixels. An explicit record of layout is the next step. For example, a snowy mountain image will have bluer regions on top, whiter regions in the middle, then a bluer region at the bottom (the lake at the foot of the mountain), whereas a waterfall image will have a darker region on the left and right and a lighter vertical stripe in the center. These layout templates can be learned for a range of images and appear to provide a significant improvement over a color histogram (Lipson et al., 1997).

Looking at image texture is a natural next step, because texture is the difference between, for example, a field of flowers (many small orange blobs), a single flower (one big orange blob), or a dalmatian and a zebra. Most people know texture when they see it, though the concept is either difficult or impossible to define. Typically, textures are thought of as spatial arrangements of small patterns—e.g., a tartan is an arrangement of small squares and lines, and the texture of a grassy field is an arrangement of thin bars.

The usual strategy for finding these subpatterns is to apply a linear filter to the image where the kernel of the filter looks similar to the pattern element. From filter theory, we have that strong responses from these filters suggest the presence of the particular pattern; several different filters can be applied, and the statistics of the responses in different places then yield a decomposition of the picture into spotty regions, barred regions, and the like (Ma & Manjunath, 1997a; Malik & Perona, 1989, 1990).

A histogram of filter responses is a first possible description of texture. For example, one might query for images with few small yellow blobs. This mechanism is used quite successfully in the Calphotos collection at Berkeley.[2] As Figures 3, 4, and 5 illustrate, a combination of color and blob queries can be used to find quite complex images. The system is described in greater detail in Carson and Ogle (1996).

Texture histograms have some problems with camera motion; as the person in Figure 2 approaches the camera, the checks on his shirt get bigger in the image. The pattern of texture responses could change quite substantially as a result. This is another manifestation of a problem we saw earlier with color histograms—i.e., the size of a region spanned by an object changes as the camera moves closer to, or further from, the object.

A strategy for minimizing the impact of this effect is to define a family of allowable transformations on the image—e.g., scaling the image by a factor in some range. We now apply each of these transformations and measure the similarity between two images as the smallest difference that can be obtained using a transformation. For example, we could scale one image by each legal factor and look for the smallest difference between color and texture histograms. In Figure 7, the earth-mover's distance allows a wide variety of transformations. Furthermore, in Rubner, Tomasi, and Guibas (1998), it has been coupled with a process for laying out images that makes the distance between images in the display reflect the dissimilarity between images in the collection. This approach allows for rapid and intuitive browsing.

The spatial layout of textures is a powerful cue. For example, in aerial images, housing developments have a fairly characteristic texture, and the layout of this texture gives cues to the region sought. In the Netra system, textures are classified into stylized families (yielding a "texture thesaurus") which are used to segment very large aerial images; this

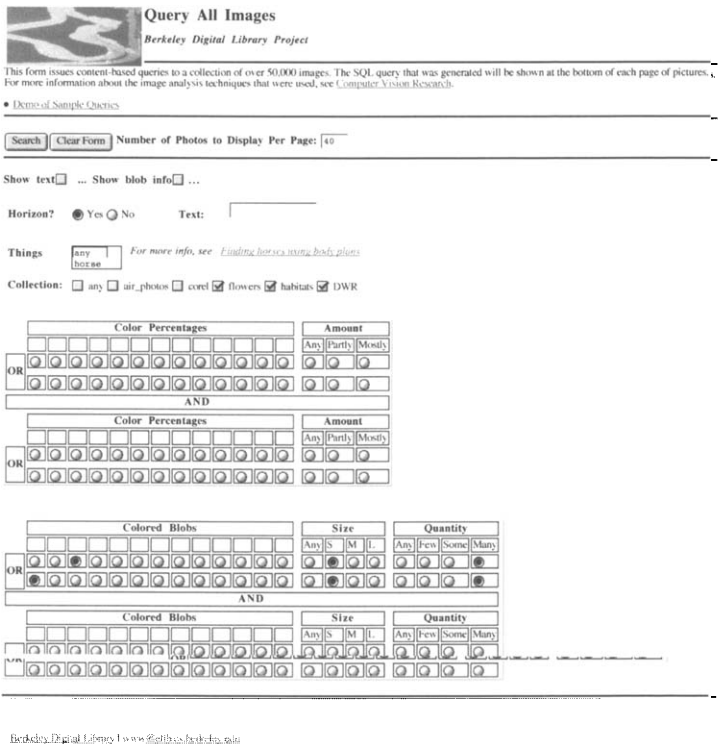


Figure 3. Specifying a query to the Calphotos collection using color and texture information. I have selected images that have a horizon and some red or yellow blobs with the intention of finding views of fields of flowers. The results appear in Figure 4.

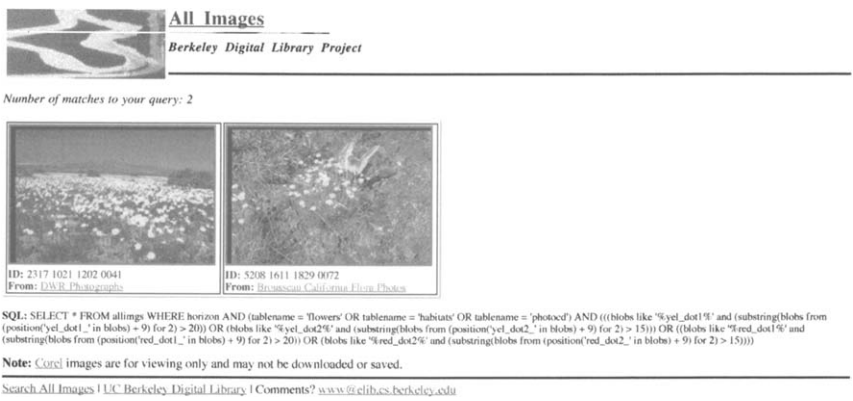


Figure 4. Images obtained from the Calphotos collection using the query from Figure 3.

Figure 5. Response images obtained by querying the Calphotos collection for images with at least one large brown blob, one or more small black blobs, and some green; this query is intended to find animals or birds.

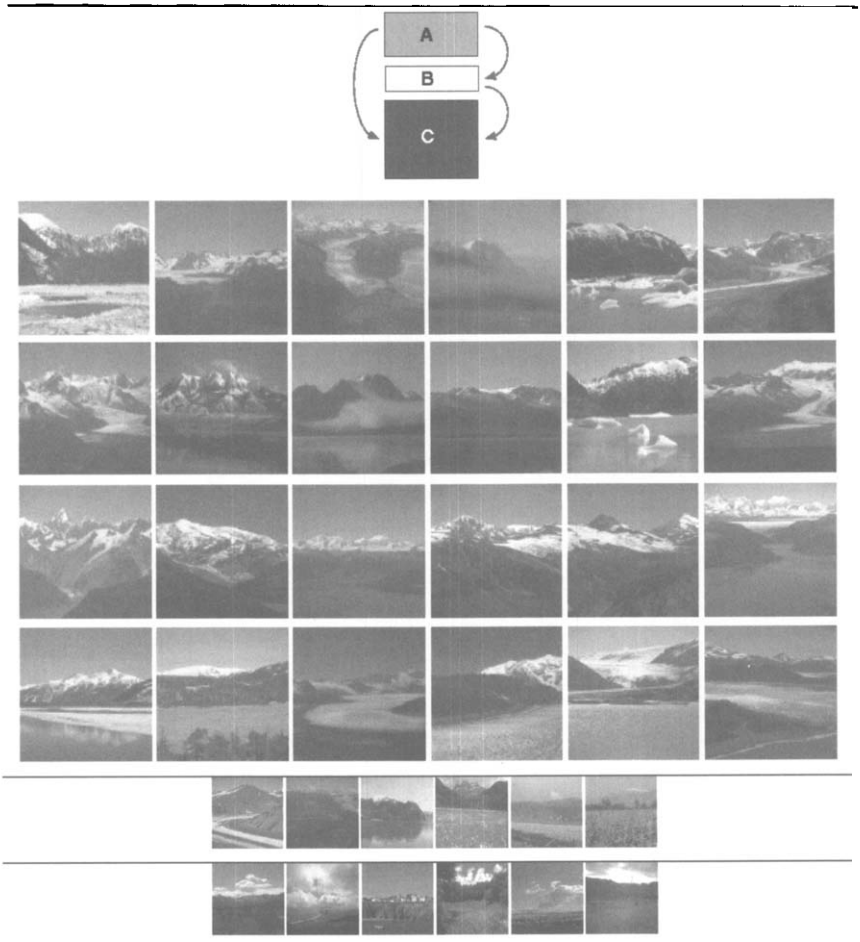


Figure 6. Spatial layout of colored regions is a natural guide to the content of many types of image. The figure on the top shows a layout of colored regions that suggests a scene showing snowy mountains; bottom center, views of mountains that were in the collection but not recovered, and bottom, images that meet the criterion but do not actually show a view of a snowy mountain. More detail appears in Lipson et al., 1997. (Figure reproduced by kind permission of W. E. L. Grimson and P. Lipson.)

Figure 7. Images laid out according to their similarity using the earth mover's distance (EMD). The EMD can be computed quickly so that displays like this—where distances between images on the display reflect the EMDs between them as faithfully as possible—can be created online. Large numbers of pictures returned from a query into an image database can thus be viewed at a glance, and a mouse click in the neighborhood of pictures that look similar to what the user is looking for tells the retrieval system where to search next. With this technology, users browse and navigate in an image database just as they would browse through a department store. Because of the large number of images displayed, and their spatially intuitive layout, users quickly form a mental model of what is in the database, and rapidly learn where to find the pictures they need. More information on the EMD can be found in Rubner et al., 1998. (Figure reproduced by kind permission of C. Tomasi.)

approach exploits the fact that, while there is a very large family of possible textures, only some texture distinctions are significant. Users can utilize example regions to query a collection for similar views—e.g., obtaining aerial pictures of a particular region at a different time or date to keep track of such matters as the progress of development, traffic patterns, or vegetation growth (see Figure 8) (Ma & Manjunath, 1997a, 1998; Manjunath, 1997a; Manjunath & Ma, 1996a, 1996b).

Regions of texture responses form patterns, too. For example, if an image shows a pedestrian in a spotted shirt, then there will be many strong responses from spot detecting filters; the region of strong responses will look roughly like a large bar. A group of pedestrians in spotted shirts will look like a family of bars, which is itself a texture. These observations suggest applying texture-finding filters to the outputs of texture-finding filters—perhaps recurring several times—and using these responses as a measure of image similarity. This strategy involves a large number of features, making it impractical to ask users to fill in a form as in Figure 9. Instead, DeBonet and Viola (1998) use an approach in which users select positive and negative example images, and the system searches for images that are similar to the positive examples and dissimilar to the negative ones.

Finding

The distinction between appearance tools and finding tools is somewhat artificial. Based on their differences in appearance, we can tell what objects are. The tools described in this section try to estimate object-level semantics more or less directly. Such systems must first segment the image—i.e., decide which pixels lie on the object of interest. Template matching systems then look for characteristic patterns associated with particular objects. Finally, correspondence reasoning can be used to identify objects using spatial relationships between parts.

Structure in a collection is helpful in finding semantics because it can be used to guide the choice of particular search mechanisms. Photobook is a system that provides three main search categories: by shape (searches for isolated objects—e.g., tools or fishes) using contour shape measured as elastic deformations of a contour; by appearance (the program can find faces using a small number of principal components); and texture (the program uses a texture representation to find textured swatches of material) (Pentland et al., 1996).

Annotation and Segmentation

A natural step in determining image semantics is to classify the type of material that image patches represent—e.g., "sky," "buildings," and so on, as opposed to "blue" or "grey." Generally, this kind of classification

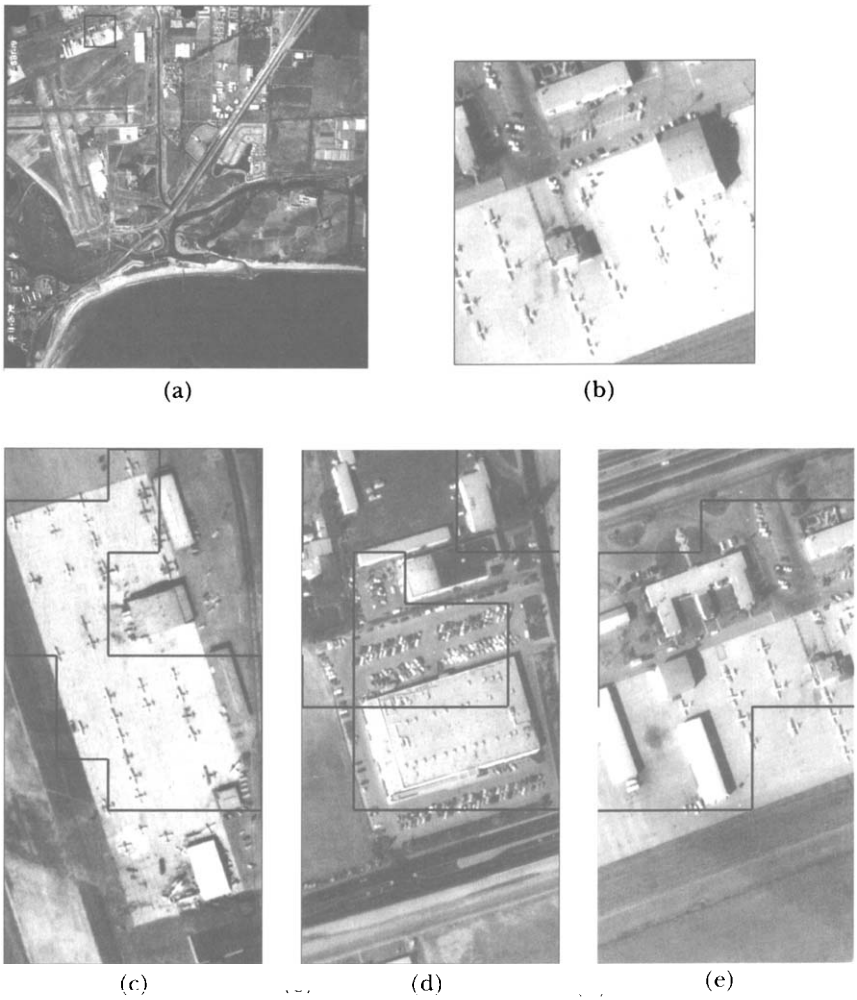


Figure 8. A texture-based search in an aerial image: (a) shows the down-sampled version of the aerial photograph from which the query is derived; (b) shows full-resolution detail of the region used for the query. The region contains aircraft, cars, and buildings. (c)-(e) show the ordered three best results of the query. Once again, the results come from three different aerial photographs. This time, the second and third results are from the same year (1972) as the query photograph but the first match is from a different year (1966). More details appear in Ma and Manjunath, 1997b. (Figure reproduced by kind permission of B. S. Manjunath.)

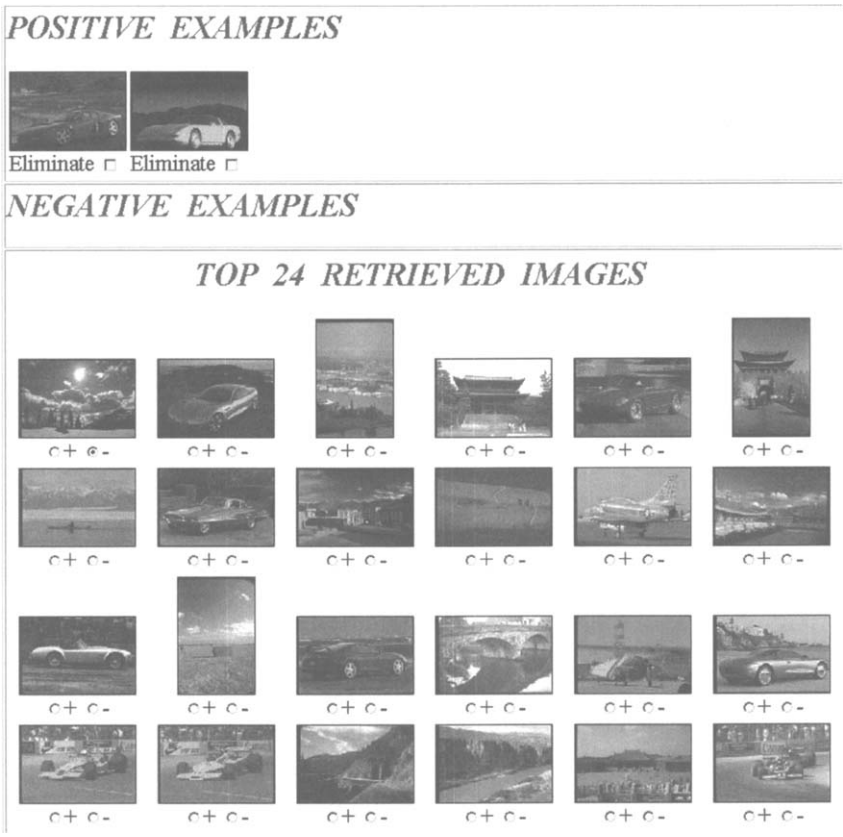


Figure 9. Querying using the "texture of textures" approach. The user has identified two pictures of cars as positive examples; these would respond strongly to large horizontal bar filters, among others. This query results in a number of returned images containing several images of cars. Figure 10 shows the effect of refining this query by exhibiting negative examples. (Figure reproduced by kind permission of P. Viola.)

Figure 10. Querying using the "texture of textures" approach. The query has been refined by providing some negative examples, yielding a response set that contains more car images. More detail on this approach appears in De Bonet & Viola, 1998. (Figure reproduced by kind permission of P. Viola.)

would need to be done with a mixture of user input (to establish appropriate categories and provide examples from those categories) and automatic annotation (for speed and efficiency). A combination of color and texture features is often, but not always, distinctive of a region; a particular difficulty is knowing which features to use and which to ignore in classifying an image patch. For example, telling sky from grass involves looking at color; telling concrete from sky may require ignoring color and emphasizing texture. Foureyes (see Figure 12) uses techniques from machine learning to infer appropriate features from user annotation practices using across-image groupings (which image patches tend to be grouped together and which apart) and in-image groupings (which patches are classified as "sky" in this image) (Minka & Picard, 1997; Picard & Minka, 1995; Minka, 1996). As a result, a user annotating an image can benefit from past experience, as illustrated in Figure 12.

Humans decompose images into parts corresponding to the objects we are interested in, and classification is one way to achieve this segmentation. Segmentation is crucial because it means that irrelevant information can be discarded in comparing images. For example, if we are searching for an image of a tiger, it should not matter whether the background is snow or grass; the tiger is the issue. However, if the whole image is used to generate measures of similarity, a tiger on grass will look very different from a tiger on snow. These observations suggest segmenting an image into regions of pixels that belong together in an appropriate sense, and then allowing the user to search on the properties of particular regions. The most natural sense in which pixels belong together is that they originate from a single object; currently, it is almost never possible to use this criterion because of an inability to know when this is the case. However, objects usually result in image regions of coherent color and texture so that pixels that belong to the same region have a good prospect of belonging to the same object.

VisualSEEK automatically breaks images into regions of coherent color and allows users to query on the spatial layout and extent of colored regions. Thus, a query for a sunset image might specify an orange background with a yellow "blob" lying on that background (Smith & Chang, 1996).

Blobworld is a system that represents images by a collection of regions of coherent color and texture (Belongie et al., 1998; Carson, Thomas, Belongie, Hellerstein, & Malik, 1999; Carson et al., 1997). The representation is displayed to the user with region color and texture displayed inside elliptical blobs that represent the shape of the image regions. The shape of these regions is represented crudely since details of the region boundaries are not cogent. A user can query the system by specifying which blobs in an example image are important and what spatial relations should hold (see Figure 11).

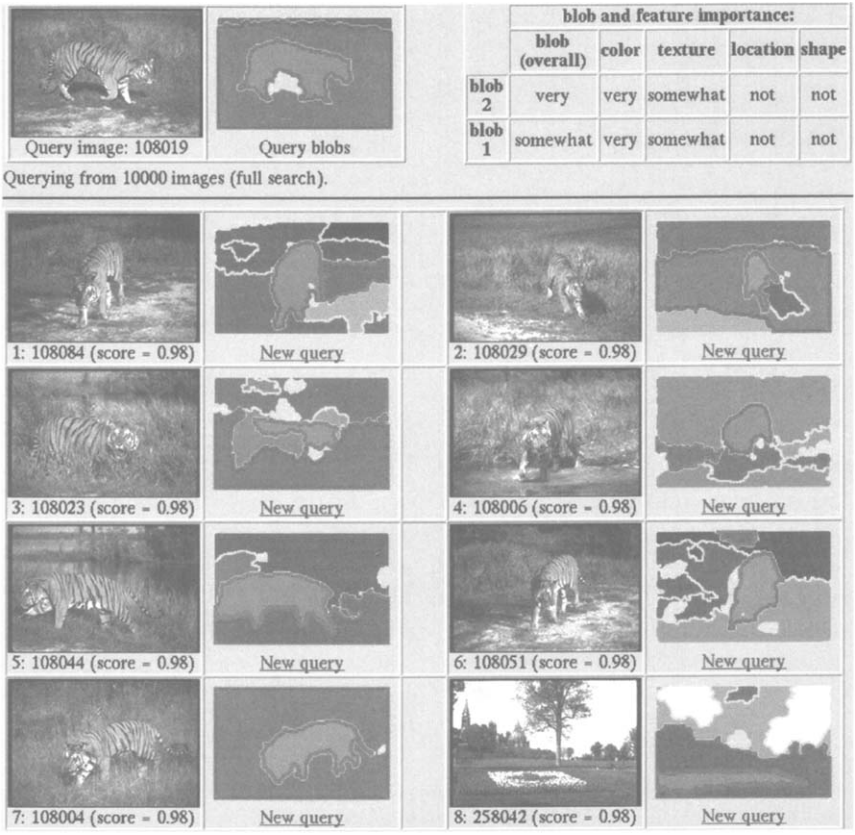


Figure 11. Blobworld query for tiger images. Users of image databases generally want to find images containing particular objects, not images with particular global statistics. The Blobworld representation facilitates such queries by representing each image as a collection of regions (or "blobs") which correspond to objects or parts of objects. The image is segmented into regions automatically, and each region's color, texture, and shape characteristics are encoded. The user constructs a query by selecting regions of interest. The Blobworld version of each retrieved image is shown, with matching regions highlighted; displaying the system's internal representation in this way makes the query results more understandable and aids the user in creating and refining queries. Experiments show that queries for distinctive objects, such as tigers and cheetahs, have much higher precision using the Blobworld system than using a similar system based only on global color and texture descriptions. Blobworld is described in greater detail in Belongie et al., 1998; Carson et al., 1999. (Figure reproduced by kind permission of C. Carson.)

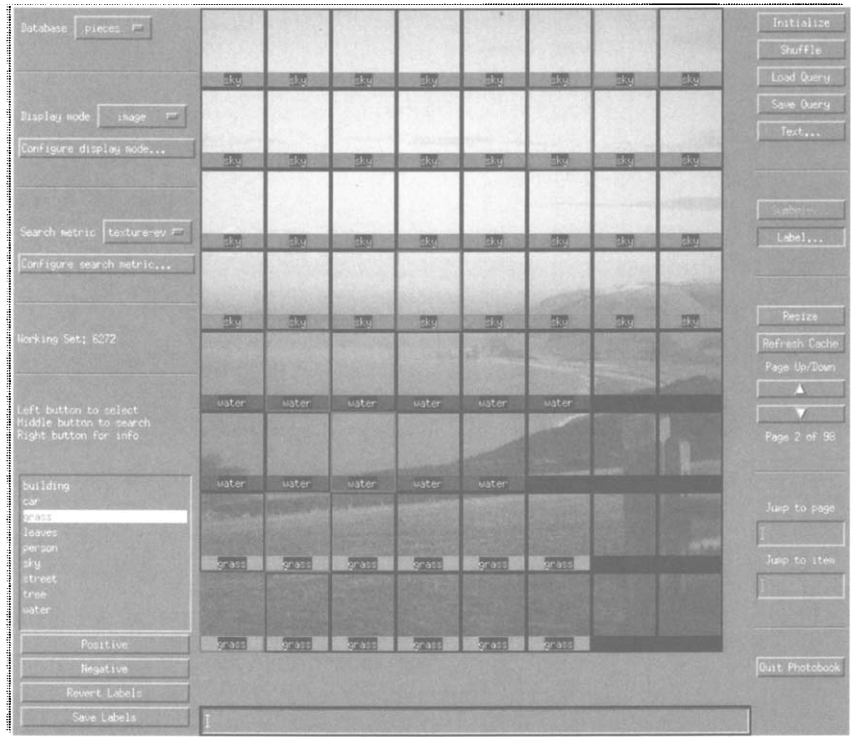


Figure 12. An image annotated with Foureyes. Red patches of image have been explicitly labeled "sky," "grass," or "water." Other labels are inferred by Foureyes from previous annotations and from the information given in these examples. More information appears in Minka and Picard, 1997. (Figure reproduced by kind permission of R. Picard.)

Template Matching

Some objects have such a distinctive appearance that they have a wide range of viewing directions and conditions. Template matching is an object recognition strategy that finds objects by matching image patches with example templates. A natural application of template matching is to construct whole-image templates that correspond to particular semantic categories (see Figure 13) (Chang et al., 1998b). These templates can be constructed offline and used to simplify querying by allowing a user to apply an existing template rather than compose a query.

Face finding is a particularly good case for template matching. Frontal views of faces are extremely similar, particularly when the face is viewed at low resolution—the main features are a dark bar at the mouth; dark blobs where the eyes are; and lighter patches at the forehead, nose, and near the mouth. This means that faces can be found, independent of the identity of the person, by looking for this pattern. Typical face finding systems extract small image windows of a fixed size, prune these windows to be oval, correct for lighting across the window, and then use a learned classifier to tell whether a face is present in the window (see Figure 14) (Rowley et al., 1996a, 1996b, 1998a; Poggio & Sung, 1995). This process works for both large and small faces because windows are extracted from images at a variety of resolutions (windows from low resolution images yield large faces and those from high resolution images yield small faces). Because the pattern changes when the face is tilted to the side, this tilt must be estimated and corrected for; this is done using a mechanism learned from data (Rowley et al., 1998b). Knowing where the faces are is extremely useful because many natural queries refer to the people present in an image or a video.

Shape and Correspondence

If object appearance can vary, template matching becomes more difficult as one is forced to adopt many more templates. There is a good template matching system for finding pedestrians, which appears to work because images of pedestrians tend to be seen at low resolution and with their arms at their sides (Oren et al., 1997). However, building a template matching system to find people is intractable because clothing and configuration can vary widely. The general strategy for dealing with this difficulty is to look for smaller templates—perhaps corresponding to "parts"—and then look for legal configurations.

One version of this technique involves finding "interest points"—points where combinations of measurements of intensity and its derivatives take on unusual values—e.g., at corners. The spatial arrangement of these points is quite distinctive in many cases. For example, as Figure 15 illustrates, the arrangement of interest points in an aerial view of Marseille is unaffected by the presence of cars; this means that one can recover and

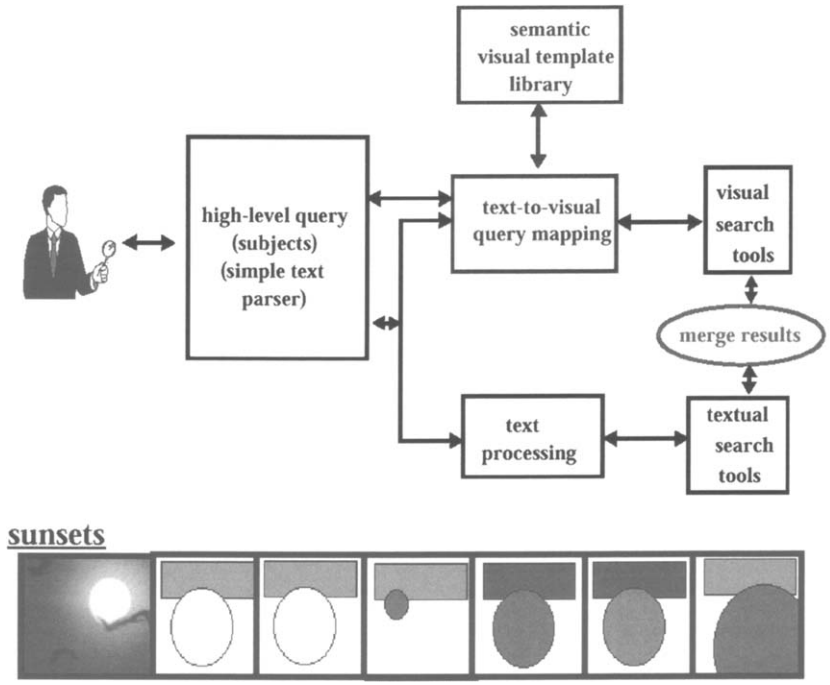


Figure 13. To remove the burden of drawing detailed low-level sketches from users, the Semantic Visual Template system helps users to develop personalized search templates. The library of semantic templates can then be used to assist users in high-level multimedia query. The top figure shows the overall structure of a system using semantic templates, and the lower figure shows a template for sunset images. More details appear in Chang et al., 1998b. (Figure reproduced by kind permission of S-F. Chang.)

Figure 14. Faces found by a learned template matching approach; the eye holes in the green mask are to indicate the orientation of the face. More details appear in Rowley et al., 1996b,1998a, 1998b. (Figure reproduced by kind permission of T. Kanade.)

register aerial images of the same region taken at different times of day using this technique. Furthermore, once interest points have been matched, an image-image transformation is known, which can be used to register the images. Registration yields further evidence to support the match and can be used to compare, for example, traffic by matching the two images at specific points.

This form of correspondence reasoning extends to matching image components with object parts at a more abstract level. People and many animals can be thought of as assemblies of cylinders (corresponding to body segments). A natural finding representation uses grouping stages to assemble image components that could correspond to appropriate body segments or other components.

This representation has been used for two cases. The first example identifies pictures containing people wearing little or no clothing. This is an interesting example: first, it is much easier than finding clothed people because skin displays very little variation in color and texture in images, whereas the appearance of clothing varies widely; second, many people are interested in avoiding or finding images based on whether they contain unclad people. This program has been tested on an unusually large and unusually diverse set of images; on a test collection of 565 images known to contain lightly clad people and 4,289 control images with widely varying content, one tuning of the program marked 241 test images and 182 control images (more detailed information appears in Forsyth et al., 1996; Forsyth & Fleck, 1996). The second example used a representation whose combinatorial structure—the order in which tests were applied—was built by hand, but the tests were learned from data. This program identified pictures containing horses and is described in greater detail in Forsyth and Fleck (1997). Tests used 100 images containing horses and 1,086 control images with widely varying content—for a typical configuration, the program marks eleven images of horses and four control images.

Video

While video represents a richer source of information than still images, the issues remain largely the same. Videos are typically segmented into shots—short sequences that contain similar content—and techniques of the form described applied within shots. Because a large change between frames is a strong cue that a shot boundary has been reached, segmenting video into shots is usually done using measures of similarity like those described in the earlier section on Appearance (e.g., Boreczky & Rowe, 1996).

The motion of individual pixels in a video is often called *optic flow* and is measured by attempting to find pixels in the next frame that correspond to a pixel in the previous frame (correspondence being measured by similarity in color, intensity, and texture). In principle, there is an

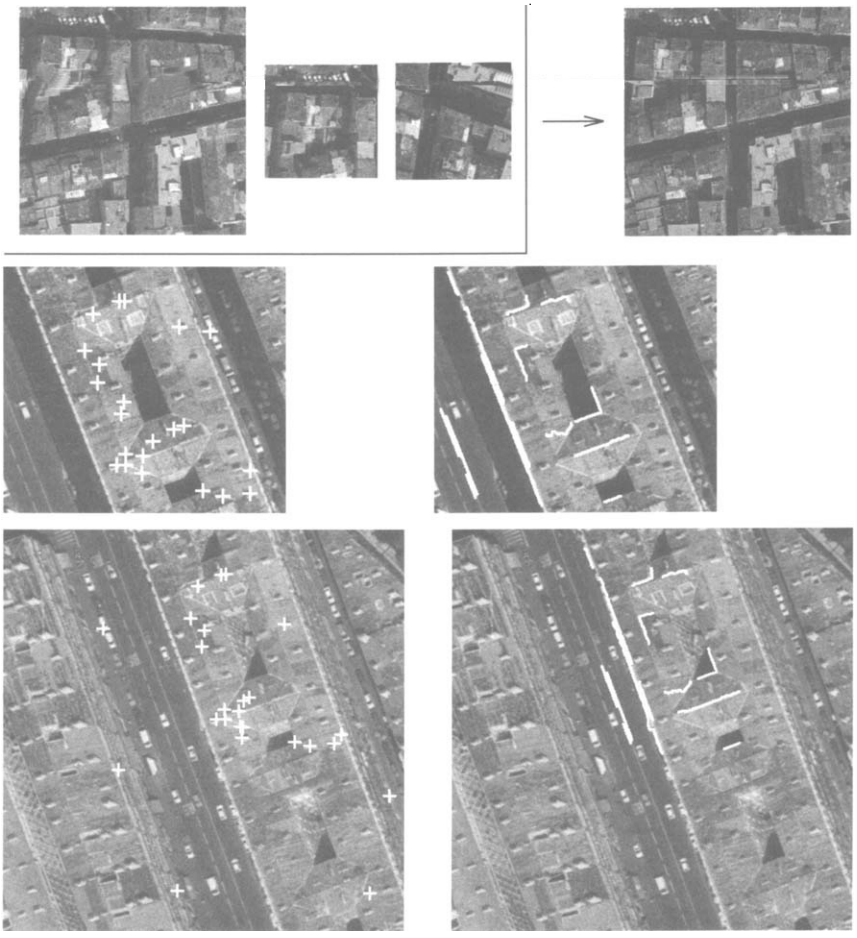


Figure 15. Images can be queried by detecting "interest points" on the image and then matching on configurations of these points based on their geometric distribution and the grey level pattern surrounding each point. The matching process is very efficient (it uses "indexing") and is tolerant of missing points in the configuration. In the example shown here, the image on the top right can be correctly retrieved from a collection of paintings, aerial images, and images of 3D objects using any of the images on the top left. Interest points used during the matching process, indicated by white crosses, for a query image (small inset on left) and the best match (bottom left). Additional evidence is obtained from the image-image transformation to confirm that the match is correct; on the right, edges which match under this transformation in the query (inset) and the result (bottom right). Notice that the two images have been taken from different viewpoints so that the building's shape differs between images. Also the scenes are not identical because cars have moved. Further details are given in Schmid and Mohr, 1997; Schmid et al., in press. (Figure reproduced by kind permission of A. Zisserman.)

optic flow vector at each pixel forming a *motion field.* In practice, it is extremely hard to measure optic flow reliably with featureless pixels because they could correspond to almost anything. For example, consider the optic flow of an egg rotating on its axis; there is very little information about what the pixels inside the boundary of the egg are doing because each looks like the other.

Motion fields can be extremely complex, however. If there are no moving objects in the frame, it is possible to classify motion fields corresponding to the camera shot used. For example, a pan shot will lead to strong lateral motion, and a zoom leads to a radial motion field. This classification is usually obtained by comparing the measured motion field with a parametric family (e.g., Sawhney & Ayer, 1996; Smith & Kanade, 1997).

Complex motion sequences are difficult to query without segmentation because much of the motion may be irrelevant to the query—e.g., in a soccer match, the motion of many players may not be significant. In VideoQ, motion sequences are segmented into moving blobs and then queried on the color and motion of a particular blob (see Figure 17) (Chang et al., 1997a; Chang et al., 1998).

The Informedia project at CMU has studied the preparation of detailed skims of video sequences. In this case, a segment of video is broken into shots, shots are annotated with the camera motion in shot, with the presence of faces, with the presence of text in shot, with keywords from the transcript, and with audio level (see Figure 18). This information yields a compact representation—the "skim"—which gives the main content of the video sequence (details in Smith & Kanade, 1997; Wactlar et al., 1996; Smith & Christel, 1995; Smith & Hauptmann, 1995).

Conclusion

For applications where the colors, textures, and layout of the image are all strongly correlated with the kind of content desired, a number of usable tools exist to find images based on content.

There has been a substantial amount of work on user interfaces and browsing, although this work is usually done to get a system up and running. Because color, texture, and layout are at best a rough guide to image content, puzzling search results are usually guaranteed. There is not yet a clear theory of how to build interfaces that minimize the impact of this effect. The most widely adopted strategy is to allow quite fluid browsing.

When queries occur at a more semantic level, we encounter deep and poorly understood problems in object recognition. Object recognition seems to require segmenting images into coherent pieces and reasoning about the relationships between those pieces. This rather vague view of recognition can be exploited to produce segmented representations that

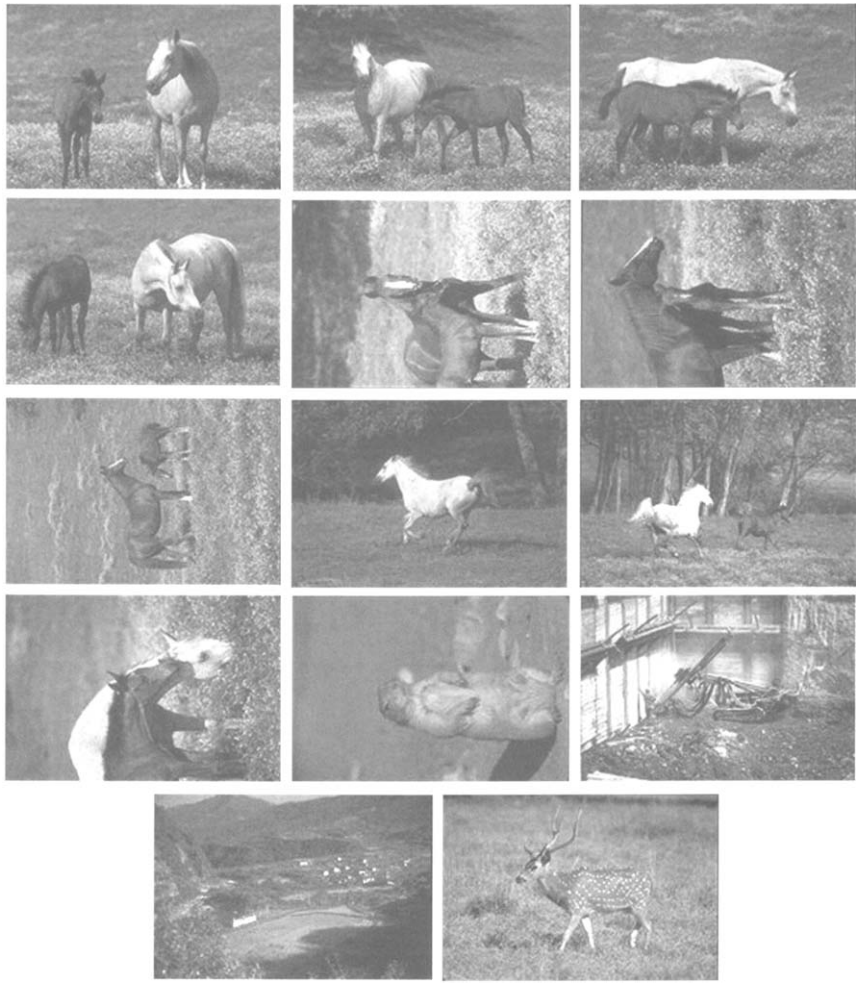


Figure 16. Images of horses recovered using a body plan representation from a test collection consisting of 100 images containing horses and 1,086 control images with widely varying content. Note that the method is relatively insensitive to aspect, but can be fooled by brown horse-shaped regions. More details appear in Forsyth and Fleck, 1997.

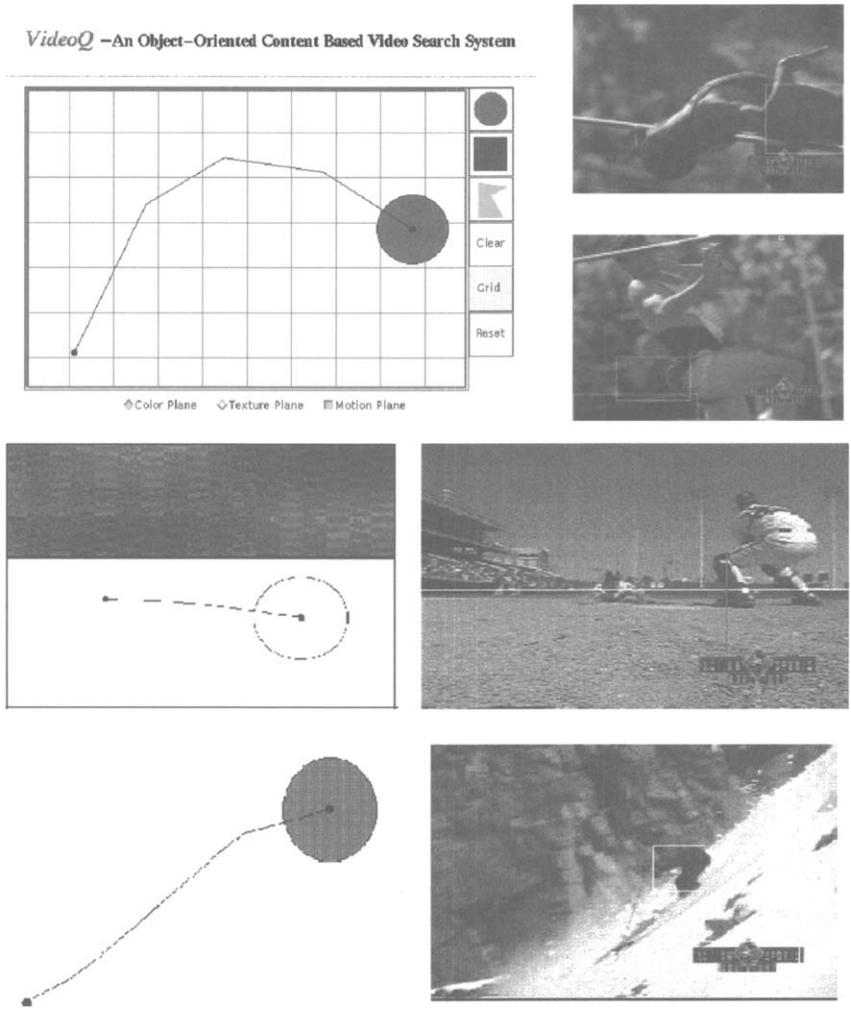


Figure 17. Video can be represented by moving blobs; sequences can then be queried by specifying blob properties and motion properties desired. The top left shows a query for a blob moving along a parabolic arc, sketched in the user interface for the VideoQ system. Top right shows frames from two sequences returned. As the center (baseball) and bottom (skiing) figures show, the mechanism extends to a range of types of motion. More details appear in Chang et al., 1997a. (Figure reproduced by kind permission of S-F. Chang.)

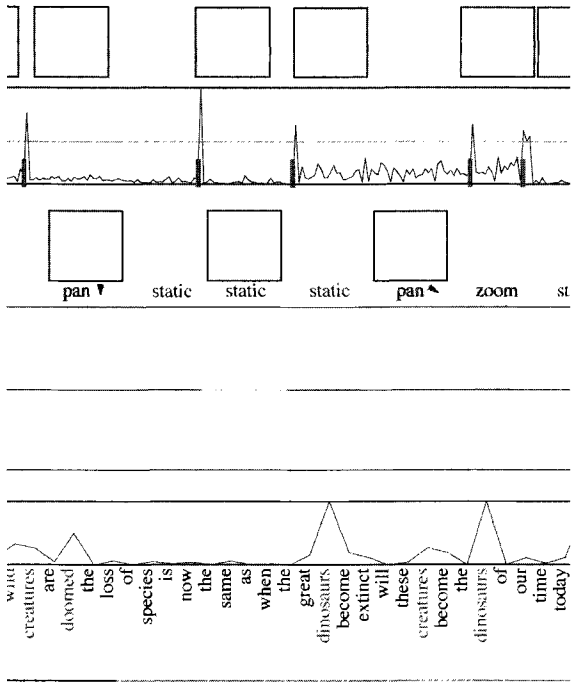


Figure 18. Characterizing a video sequence to create a skim. The video is segmented into scenes. Camera motions are detected along with significant objects (faces and text). Bars indicate frames with positive results. Word relevance is evaluated in the transcript. More information appears in Smith and Kanade, 1997. (Figure reproduced by kind permission of T. Kanade.)

allow searches for objects independent of their backgrounds. Furthermore, some special cases of object recognition can be handled explicitly. It is not known how to build a system that could search for a wide variety of objects; building a user interface for such a system would present substantial problems, too. Some images have text associated with them, either because content providers have explicitly described the image or because images can be associated with captions or document text. In this case, it is natural to use these terms to obtain semantic descriptions. There is currently no clear set of principles to use to combine available text descriptions with images. What lines of research are more promising? The answer depends on what application one has in mind. Some applications—e.g., reviewing images or video in libraries to choose a good background picture for an advertisement—really need images to be arranged by their appearance so that all the soothing blue pictures are in about the same place. Other applications demand semantic descriptions that are very difficult to supply; there is no way to answer a question like "show trends in the pose of subject in portraits in the seventeenth century" without well developed processes for finding people in pictures. General semantic descriptions are a long way off, but one natural focus is the activities of people. I expect that, in the next few years, using entirely automatic methods, it may be possible to find pictures of, for example, a politician kissing a baby.

Acknowledgments

Many people kindly allowed me to use their figures and suggested captions. Thanks to: Shih-Fu Chang of Columbia; Takeo Kanade, Henry Rowley, and Michael Smith of CMU; Jitendra Malik and Chad Carson of U.C. Berkeley; B.S. Manjunath of U.C. Santa Barbara; Roz Picard of the MIT Media Lab; Cordelia Schmid of INRIA; Carlo Tomasi and Leo Guibas of Stanford; Paul Viola of the MIT AI Lab; Ramin Zabih of Cornell; and Andrew Zisserman of Oxford University. Margaret Fleck read a draft under trying circumstances. Portions of the research described in this article were supported by a Digital Library grant (NSF-IRI-9411334).

Notes

[1] A collection of 60,000 images quite commonly used in vision research available in three series from the Corel corporation, whose head office is at 1600 Carling Avenue, Ottawa, Ontario, K1Z 8R7, Canada.

[2] At http://elib.cs.berkeley.edu/photos, there are many thousands of images of California natural resources, flowers, and wildlife.

References

Armitage, L. H., & Enser, P. G. B. (1997). Analysis of user need in image archives. *Journal of Information Science, 23*(4), 287-299.

Belongie, S.; Carson, C.; Greenspan, H.; & Malik, J. (1998). Color and texture-based image segmentation using EM and its application to content based image retrieval. In *Proceedings of the IEEE 6th International Conference on Computer Vision* (Bombay, India, January 4-7, 1998) (pp. 675-682). New Delhi, India: Narosa Publishing House.

Boreczky, J. S., & Rowe, L. A. (1996). Comparison of video shot boundary detection

techniques. *Journal of Electronic Imaging, 5*(2), 122-128.

Carson, C., & Ogle, V. E. (1996). Storage and retrieval of feature data for a very large online image collection. In S. Y. W. Su (Ed.), *Proceedings of the 12th International Conference on data engineering.* Los Alamitos, CA: IEEE Computer Society.

Carson, C.; Thomas, M.; Belongie, S.; Hellerstein, J. M.; & Malik, J. (1999). Blobworld: A system for region-based image indexing and retrieval. In D. P. Huijsmans & A. W. M. Smeulders (Eds.), *Visual information systems* (Proceedings of the 3rd International Conference, Visual '99, June 2-4, 1999) (pp. 509-516). New York: Springer.

Chang, S.-F.; Chen, W.; Meng, H. J.; Sundaram, H.; & Zhong, D. (1997a). VideoQ: An automatic content-based video search system using visual cues. In *Proceedings of the Fifth ACM International Multimedia Conference* (November 9-13, 1997, Seattle, WA) (pp. 313-324). New York: Association for Computing Machinery Press.

Chang, S.-F.; Smith, J. R.; Beigi, M.; & Benitez, A. (1997b). Visual information retrieval from large distributed online repositories. *Communications of the ACM, 40*(12), 63-71.

Chang, S.-F.; Chen, W.; Meng, H. J.; Sundaram, H.; & Zhong, D. (1998a). A fully automated content based video search engine supporting spatiotemporal queries. *IEEE Transactions on Circuits & Systems for Video Technology, 8*(8), 602-615.

Chang, S.-F.; Chen, W.; & Sundaram, H. (1998b). In *ICIP '98: Proceedings of the 1998 International Conference on Image Processing* (October 4-7, 1998, Chicago, IL). Los Alamitos, CA: IEEE Computer Society.

Chapelle, O.; Haffner, P.; & Vapnik, V. (1999). Support vectors for histogram-based classification. *IEEE Transactions on Neural Networks, 10*(5), 1055-1064.

Congiu, G.; Del Bimbo, A.; & Vicario, E. (1995). Iconic retrieval by contents from databases of cardiological sequences. In *Visual database systems 3: Visual information management* (Proceedings of the 3rd IFIP 2.6 Working Conference on Visual Database Systems, March 27-29, 1995, Lausanne, Switzerland) (pp. 158-174). London: Chapman & Hall.

De Bonet, J. S., & Viola, P. (1998). Structure driven image database retrieval. In M. I. Jordan, M. J. Kearns, & S. A. Solla (Eds.), *Advances in neural information processing systems* (vol. 10, pp. 866-872). Cambridge, MA: MIT.

Eakins, J.; Boardman, P.; & Graham, M. E. (1998). Similarity retrieval of trademark images. *IEEE Multimedia, 5*(2), 53-63.

Enser, P. G. B. (1993). Query analysis in a visual information retrieval context. *Journal of Document and Text Management, 1*(1), 25-52.

Enser, P. G. B. (1995). Pictorial information retrieval. *Journal of Documentation, 51*(2), 126-170.

Flickner, M.; Sawhney, H.; Niblack, W.; Ashley, J.; Huang, Q.; Dom, B.; Gorkani, M.; Hafner, J.; Lee, D.; Petkovic, D.; & Steele, D. (1996). Query by image and video content: The QBIC system. *Computer, 28*(9), 23-32.

Forsyth, D. A., & Fleck, M. M. (1996). Identifying nude pictures. In *Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision, WACV '96* (December 2-6, 1996, Sarasota, FL) (pp. 103-108). Los Alamitos, CA: IEEE Computer Society.

Forsyth, D. A., & Fleck, M. M. (1997). Body plans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (San Juan, PR) (pp. 678-683). Los Alamitos, CA: IEEE Computer Society.

Forsyth, D. A.; Fleck, M. M.; & Bregler, C. (1996). Finding naked people. In B. Buxton & R. Cipollo (Eds.), *Computer Vision, ECCV '96* (Proceedings of the 4th European Conference on Computer Vision, Cambridge, United Kingdom, April 14-18, 1996) (pp. 593-602). Berlin, Germany: Springer-Verlag.

Forsyth, D. A., & Ponce, J. (in press). *Computer vision: A modern approach.* Upper Saddle River, NJ: Prentice-Hall.

Hampapur, A.; Gupta, A.; Horowitz, B.; Shu, C.-F.; Fuller, C.; Bach, J.; Gorkani, M.; & Jain, R. (1997). Virage video engine. In *Storage and retrieval for image and video databases V* (Proceedings of SPIE, The International Society for Optical Engineering) (vol. 3022, pp. 188-198). Bellingham, WA: SPIE.

Holt, B., & Hartwick, L. (1994a). Quick, who painted fish?: Searching a picture database with the QBIC project at UC Davis. *Information Services and Use, 14*(2), 79-90.

Holt, B., & Hartwick, L. (1994b). Retrieving art images by image content: The UC Davis QBIC project. *ASLIB Proceedings, 46*(10), 243-248.

Huang, J., & Zabih, R. (1998). *Combining color and spatial information for content-based image retrieval.* Retrieved July 12, 1999 from the World Wide Web: http://www.cs.cornell.edu/html/rdz/Papers/ECDL2/spatial.htm.

Jacobs, C. E.; Finkelstein, A.; & Salesin, D. H. (1995). Fast multiresolution image querying. In *Proceedings of SIGGRAPH '95* (August 6-11,1995, Los Angeles, CA) (pp. 277-285). New York: Association of Computing Machinery Press.

Jain, A. K., & Vailaya, A. (1998). Shape-based retrieval: A case study with trademark image databases. *Pattern Recognition, 31*(9), 1369-1390.

La Cascia, M.; Sethi, S.; & Sclaroff, S. (1998). Combining textual and visual cues for content based image retrieval on the World Wide Web. In *IEEE workshop on content based access of image and video libraries* (pp. 24-28). Los Alamitos, CA: IEEE Computer Society.

Lipson, P.; Grimson, W. E. L.; & Sinha, P. (1997). Configuration based scene classification and image indexing. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (June 17-19, 1997, San Juan, PR) (pp. 1007-1013). Los Alamitos, CA: IEEE Computer Society.

Ma, W. Y., & Manjunath, B. S. (1997a). Edge flow: A framework for boundary detection and image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (June 17-19, 1997, San Juan, PR) (pp. 744-749). Los Alamitos, CA: IEEE Computer Society.

Ma, W. Y., & Manjunath, B. S. (1997b). NeTra: A toolbox for navigating large image databases. In *Proceedings of the IEEE international conference on image processing* (October 26-29, 1997, Santa Barbara, CA) (pp. 568-571). Los Alamitos, CA: IEEE Computer Society.

Ma, W. Y., & Manjunath, B. S. (1998). A texture thesaurus for browsing large aerial photographs. *Journal of the American Society for Information Science, 49*(7), 633-648.

Malik, J., & Perona, P. (1989). A computational model of texture segmentation. In *Proceedings of the 22nd Asilomar Conference on Signals, Systems, and Computers* (October 31-November 2, 1988, Pacific Grove, CA, Naval Postgraduate School, San Jose State University) (pp. 490-494). San Jose, CA: Maple Press.

Malik, J., & Perona, P. (1990). Preattentive texture discrimination with early visual mechanisms. *Journal of the Optical Society of America: A-Optics & Image Science,* 7(5), 923-932.

Manjunath, B. S., & Ma, W. Y. (1996a). Browsing large satellite and aerial photographs. In In *Proceedings of the 3rd IEEE international conference on image processing* (September 16-19, 1996, Lausanne, Switzerland) (pp. 765-768). New York: IEEE Computer Society.

Minka, T. P., & Picard, R. W. (1997). Interactive learning with a "society of models." *Pattern Recognition, 30*(4), 565-581.

Minka, T. P. (1996). *An image database browser that learns from user interaction* (MIT Media Laboratory Perceptual Computing Section Tech. Rep. No. 365). Cambridge, MA: MIT.

Mundy, J. L., & Vrobel, P. (1994). The role of IU technology in radius phase II. In *Proceedings of the 23rd Image Understanding Workshop* (November 13-16, 1994, Monterey, CA) (pp. 251-264). San Francisco: Morgan Kaufmann.

Mundy, J. L. (1995). The image understanding environment program. *IEEE Intelligent Systems and their Applications, 10*(6), 64-73.

Mundy, J. L. (1997). IU for military and intelligence applications: How automatic will it get? In *Emerging applications of computer vision* (Proceedings of SPIE, the Society for Optical Engineering, vol. 2962) (pp. 162-170). Bellingham, WA: SPIE.

Ogle, V. E., & Stonebraker, M. (1995). Chabot: Retrieval from a relational database of images. *Computer, 28*(9), 40-48.

Oren, M.; Papageorgiou, C.; Sinha, P.; & Osuna, E. (1997). Pedestrian detections using wavelet templates. In *IEEE Computer Society conference on computer vision and pattern recognition* (June 17-19, 1997, San Juan, PR) (pp. 193-199). Los Alamitos, CA: IEEE Computer Society.

Pentland, A.; Picard, R.; & Sclaroff, S. (1996). Photobook: Content-based manipulation of image databases. *International Journal of Computer Vision, 18*(3), 233-254.

Picard, R. W., & Minka, T. (1995). Vision texture for annotation. *Journal of Multimedia Systems, 3*(1), 3-14.

Poggio, T., & Sung, K.-K. (1995). Finding human faces with a gaussian mixture distribution-based model. In *AACV '95* (Proceedings of the 2nd Asian Conference on Com-

puter Vision) (pp. 435-440). Singapore: Nanyang Technological University.

Psarrou, A.; Konstantinou, V.; Morse, P.; & O'Reilly, P. (1997). Content based search in medieval manuscripts. In *TENCON '97* (Proceedings of the IEEE TENCON '97, IEEE Region 10 Annual Conference: Speech and image technologies for computing and telecommunications (December 2-4, 1997, Queensland University of Technology, Brisbane, Australia) (pp. 187-190). New York: IEEE Computer Society.

Rowley, H. A.; Baluja, S.; & Kanade, T. (1996a). Human face detection in visual scenes. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing 8* (Proceedings of the 1995 conference) (pp. 875-881). Cambridge, MA: MIT.

Rowley, H. A.; Baluja, S.; & Kanade, T. (1996b). Neural network based face detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (June 18-20, 1996, San Francisco, CA) (pp. 203-208). Los Alamitos, CA: IEEE Computer Society.

Rowley, H. A.; Baluja, S.; & Kanade, T. (1998a). Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(1), 23-38.

Rowley, H. A.; Baluja, S.; & Kanade, T. (1998b). Rotation invariant neural network-based face detection. In *Proceedings of the 1998 IEEE conference on computer vision and pattern recognition* (June 23-25, 1998, Santa Barbara, CA) (pp. 38-44). Los Alamitos, CA: IEEE Computer Society.

Rubner, Y.; Tomasi, C.; & Guibas, L. J. (1998). A metric for distributions with applications to image databases. In *Proceedings of the 6th international conference on computer vision* (January 4-7, 1998, Bombay, India) (pp. 59-66). New Delhi, India: Narosa Publishing House.

Sawhney, H., & Ayer, S. (1996). Compact representations of videos through dominant and multiple motion estimation. *IEEE Transactions on Analysis and Machine Intelligence, 18*(8), 814-830.

Schmid, C., & Mohr, R. (1997). Local gray value invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(5), 530-534.

Schmid, C.; Zisserman, A.; & Mohr, R. (in press). Integrating geometric and photometric information for image retrieval. In *International workshop on shape, contour, and grouping in computer vision.*

Seloff, G. A. (1990). Automated access to the NASA-JSC image archives. *Library Trends, 38*(4), 682-696.

Smith, J. R., & Chang, S.-F. (1996). VisualSEEK: A fully automated content-based image query system. In *Proceeding of ACM multimedia '96* (November 18-22, 1996, Boston, MA) (pp. 87-98). New York: Association for Computing Machinery Press.

Smith, J. R., & Chang, S.-F. (1997). Visually searching the Web for content. *IEEE Multimedia, 4*(3), 12-20.

Smith, M. A., & Christel, M. G. (1995). Automating the creation of a digital video library. In *Proceedings of ACM multimedia '95* (November 5-9, 1995, San Francisco, CA) (pp. 357-358). New York: Association for Computing Machinery Press.

Smith, M. A., & Hauptmann, A. (1995). Text, speech and vision for video segmentation: The informedia project. In *AAAI Fall 1995 symposium on computational models for integrating language and vision.* Menlo Park, CA: AAAI Press.

Smith, M., & Kanade, T. (1997). Video skimming for quick browsing based on audio and image characterization. In *1997 IEEE computer society conference on computer vision and pattern recognition* (June 17-19, 1997, San Juan, PR). Los Alamitos, CA: IEEE Computer Society.

Smith, T. R. (1996). A digital library for geographically referenced materials. *Computer, 29*(5), 54-60.

Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision, 7*(1), 11-32.

Trucco, E., & Verri, A. (1998). *Introductory techniques for 3-D computer vision.* Upper Saddle River, NJ: Prentice-Hall.

Wactlar, H.; Kanade, T.; Smith, M.; & Stevens, S. (1996). Intelligent access to digital video: The Informedia project. *Computer, 29*(5), 46-52.

Wong, S. T. C. (1998). CBIR in medicine: Still a long way to go. In *Proceedings of the workshop on content-based access of image and video libraries* (June 21, 1998, Santa Barbara, CA) (p. 114). Los Alamitos, CA: IEEE Computer Society.

Implementation and Evaluation

"Securing Digital Image Assets in Museums and Libraries: A Risk Management Approach," *Teresa Grose Beamsley*

"Getting the Picture: Observations from the Library of Congress on Providing Online Access to Pictorial Images," *Caroline R. Arms*

"Recent Developments in Cultural Heritage Image Databases: Directions for User-Centered Design," *Christie R. Stephenson*

"Evaluation of Image Retrieval Systems: Role of User Feedback," *Samantha K. Hastings*

Securing Digital Image Assets in Museums and Libraries: A Risk Management Approach

Teresa Grose Beamsley

Abstract

There is an obvious need for ongoing research, evaluation, and planning if museums and archives are committed to protecting their digital image assets. A number of potential threats to the integrity of digital image information can be identified when standard practices in museums and archives are examined. Changes in the integrity of digital image information can be caused by the manner in which the source data are acquired and recorded and by modifications made to the image data file. Alterations made to contextual data can limit valid interpretation of the associated surrogate image. The destruction of the mechanisms that link contextual data to the appropriate digital image has the same effect as deleting contextual information. Loss of control over digital assets can be the result of failure or inability to establish and publicize copyright. Even if copyright is established and enforceable, failure to enforce rights has the same effect as having no rights at all. Finally, failure to detect corruption of digital information means that invalid, partial, or inappropriate information will be spread under the guise of authentic reliable information.

Some institutions are already proactively applying security measures to digital image collections. Some of these security measures can have a negative impact on the integrity of the files that they are designed to protect. Systematic consideration of risk factors can inform the creation of procedures and application of security that works to guarantee the reliability and accuracy of digital image assets.

Teresa Grose Beamsley, Collections Information Resources, Henry Ford Museum & Greenfield Village, P.O. Box 1970, Dearborn MI 48121-1970
LIBRARY TRENDS, Vol. 48, No. 2, Fall 1999, pp. 359-378

Digital Image Information as an Institutional Asset

In their earliest manifestations, museums and archives were essentially collections of primary source materials. The collectors determined the criteria by which artifacts or manuscripts were chosen for preservation. The criteria were based at least in part on the value of the information that was embodied in the content of the materials or implied in the existence of the objects, a value that was established by the needs and interests of wealthy collectors.

Public exposure to museum and archival collections began in earnest at the turn of the nineteenth century. The infrequent opening of personal holdings to scrutiny became more commonplace as the general population came to recognize the existence of these collections and to demand access. In some cases, the profit motives of collection holders played a significant role in the growing accessibility of collections. The public saw value in the experience of gaining physical access to rare and unusual materials. The collectors saw value in offering access (sometimes for profit) to a new market. Selection of materials and the determination of their intrinsic information value were still determined by gentleman collectors. Increasingly, scholars used the information in their studies and augmented the utility of the collections by adding to the body of contextual data about them.

Academic research played an important role in the evolution of the modern nonprofit museum in the early twentieth century. Scholars and connoisseurs formed the basis of a class of professional museum workers. Curators, preparators, and conservators adopted codes of ethics and standards of practice that were instrumental in the development of museums and archives as educational institutions. However, until the 1950s, the primary audiences of both types of institutions were on-site visitors with specific, and often specialist, research needs rather than the casually curious.

During the past twenty years, a combination of changing professional attitudes, the interests of public and private funders, and the growing availability and reliability of reproduction technologies and electronic communication have resulted in a re-evaluation of museum and archival collections. The new target audience is the general education market and the new means of providing information to the target audience is electronic, most often via the Internet. The World Wide Web allows easy access to good quality image representations as well as to text-based contextual information about them. The public's expectation is that a broad range of information needs can and will be accomplished accurately via electronic surrogates without physical exposure to the primary sources from any place at any time. The worth of institutional assets is no longer gauged by looking at the collections inventory appraisal. It is now redefined as the combination of the physical materials in the collections, the

surrogates that satisfy a growing demand for visual information about them, and the text-based information that establishes their context and serves as the key to locating them.

Securing collections assets against misuse, theft, or damage is an ongoing concern of museums and archives. A variety of measures are implemented to safeguard collections. These include controlled access to storage and items on display, frequent inventories, environmental monitoring, administration of rights and releases, and strict procedures regarding use by staff members and others. Posting extra guards does not help to secure electronic information. And, unlike the Impressionist painting that is kept under surveillance or the Stradivarius violin that is rarely, if ever, removed from the display case, digital assets can be adversely affected by the very measures that are intended to ensure their integrity and authenticity. Security measures typically used in museums and archives to protect these assets are applied randomly at best and unintentionally at worst. Responsible stewardship of digital image assets calls for a more formal and thorough risk management assessment of potential threats and for the creation of an informed and thoughtful security plan for their management and protection.

Risk management is the sum of all activities directed toward acceptably accommodating the possibility of failure in a program. Risk management is based on assessment; every risk management assessment includes a number of tasks: (1) identification of concerns, (2) identification of risks, (3) evaluation of the risks as to likelihood and consequences, (4) assessment of options for accommodating the risks, (5) prioritization of risk management efforts, and (6) development of risk management plans (http://www.airtime.co.uk/users/wysywig/risk_1.htm). This article examines existing practices within museums and archives and provides suggestions on the creation of such plans as they apply specifically to the stewardship of digital assets.

Defining Concerns and Identifying Risks

Responsible individuals become concerned when a valuable possession is placed in jeopardy. The value of collections-related digital assets to museums and archives has been established. What are legitimate concerns regarding objects of value? Would these concerns be applicable to digital assets? It is possible to identify two obvious concerns. The first is fear that the asset itself will somehow lose value. The second is that the steward (in this case the institution and its professional staff) will somehow lose the asset or control over the asset.

How is value embodied in digital information and what would constitute a loss of value? The charter of museums and archives includes a mandate to preserve the information embodied in their collections. It seems reasonable to propose that the value of digital surrogates for

collections items lies in the relative ability of the surrogates to convey as much original information content as possible. The integrity of the digital image is judged as the degree to which it accurately represents its subject. If the information content of the surrogate is compromised, the surrogate is devalued.

There is a case to be made for the creation of very high quality, very high-resolution digital surrogates. These files are used as archival versions of image information, but reality intervenes when their content is put to practical use. High quality, high-resolution files are very large and therefore costly to store and transmit. The generally accepted rule is that the needs of different uses and users are best met by digital content presented in a variety of formats or resolutions, tailored to the situation. Accurate representation is in the eye of the beholder; the resolution and file size limitations dictated by intended Web use are not the same as those demanded by activities such as conservation assessment (Frey, 1997) (http://lcweb2.loc.gov/ammem/formats.html). As a result, every variant form of a digital file can and should be evaluated for integrity based on the use to which it is put.

Control over the asset is somewhat easier to describe and evaluate. The most obvious manifestation of control of image surrogates is the ownership of copyright and the ability to assign or to withhold assignment of use rights to others. There are other manifestations of control that are uniquely related to the museum or archive's responsibilities toward the public; these may in fact be more significant than copyright ownership. Nonprofit 501(c)3 charters and ethical responsibility dictate that it is not enough for institutions to own and care for objects. The legal definition of a museum includes the directive "to exhibit to the public on a regular basis" (Malaro, 1985). This has been interpreted for the last two decades as a mandate to educate by providing members of the public with meaningful and useful mediated access to collections. Control of the collections implies control of access to the collections in a proactive way. It is the job of museums to encourage and facilitate the use of collections and the information that they represent. Loss of control in this sense would mean an inability to effectively mediate the collections-related educational experience.

It is now possible to identify potential risks that are associated with each type of concern. Changes in the integrity of digital image information can be caused by direct modifications made to the image data. They may also be associated with modifications to contextual data that limit understanding and interpretation of the associated surrogate image. The destruction of the mechanisms that link contextual data to the appropriate digital image has the same effect as deleting contextual information. Loss of control over digital assets can be the result of failure to establish ownership and/or copyright. Even if copyright is established and

enforceable, failure to enforce rights has the same effect as having no rights at all. Failure to detect corruption of digital information means that invalid, partial, or inappropriate information will be spread under the guise of authentic reliable information. Each of these risks represents the possibility of an information systems failure.

Prioritizing Risks—How Safe is Standard Practice?

What are the chances that any of these risks will be realized? An examination of the typical ways in which digital image information and associated contextual data are created, managed, and made accessible sheds light on the probability of content degradation. Most institutions already employ both active and passive measures to prevent or minimize the impact of a reduction in reliable content in systems that depend on the use of digital image information. Do these efforts have any effect on the immediacy of each risk?

Creating Digital Image Files and Derivatives

A digital image cannot be a better representation than the best available from the conversion method used to create the image. A number of authors and research groups have conducted comparative studies of conversion techniques and produced recommendations for best-practice conversion methods, ranging from direct digital photography through microfilm and negative scanning to direct positive scanning and PhotoCD processing (http://www.columbia.edu/acis/dl/imagespec.html) (Kenney, 1997; Conway, 1996; Reilly, 1995). Similarly, the digitized image cannot be better than the source document or object without some sort of data modification. It is not necessary to belabor the importance of informed decision making in the process of creating archival image files from which derivative files may be drawn. Frey (1997) suggests that four targets be used for objectively evaluating the results of digitization: tone reproduction, color reproduction, detail and edge reproduction, and noise. Satisfactory performance in output tests of all four targets will guarantee that, at least at the archival level, an acceptably accurate digital representation of image information has been created.

The integrity of digital image information is inherent in the structure of the image file. Only bit-mapped images (those created from aggregations of discrete bits or units of data) are considered in this discussion; vector image data are created and used in museum and archival environments much less frequently than in academic libraries and special collections. The parameters that are chosen to define file structure determine the limitations of the file as an image surrogate. These parameters include dynamic range, resolution, and compression (Besser et al., 1995).

Dynamic range defines the ability of the file structure to convey tonal information about each pixel captured. Every digital image is composed

of a fixed number of pixels—tiny discrete blocks of tone. Bi-tonal images can only convey information in black and white. A bit (the basic building block of digital information) can only convey two possible values; therefore, bi-tonal information is conveyed using one bit per pixel. This type of information encoding produces the smallest possible files, but the resulting image cannot represent any range of shades between black and white. It is recommended for uses that involve modern printed works and line drawings or graphics and is frequently employed when the desired use is a printed reproduction of such materials. Gray scale uses 8 bits to represent each pixel, providing the capability of representing up to 256 shades ranging from pure white through gray to pure black. This format is usually recommended for representing black and white photographs, half-tone illustrations, and other two-dimensional representations that convey shading or variation in ink density. Color is best represented using 24 bits per pixel, which provides about 16 million different colors but which results in much larger file sizes. Color conveys much more information than gray-scale or bi-tonal files and is required for images in which color must be maintained but is also recommended for use in digitizing images of older documents (http://www.columbia.edu/acis/dl/imagespec.html; http://lcweb2.loc.gov/ammem/pictel/index.html). While software, printing, and display hardware designs determine the nature of the end product, the dynamic range of the image file establishes the bases from which these devices perform in tests of tone and color reproduction. Recording image data in a file structure that uses 8-bit color, for example, will in most cases result in image information that offers only a general approximation of the tonality of the original and severely impact the utility of the image surrogate for many uses.

Many institutions have chosen to protect their digital image assets by providing general access to only low-resolution files. Resolution refers to the number of pixels that are used to describe a single image (the fixed number mentioned earlier); it is usually expressed in terms of horizontal and vertical dimensions. An image recorded at a resolution of 512 x 768, for instance, has 512 rows and 768 columns of pixels. Resolution affects the level of detail that can be depicted by the image file. If a lower resolution is specified, fewer pixels will be used to describe the image and therefore edges may be blurred, areas of the displayed or printed image may appear blocky, tonal transitions may seem more abrupt, and detail may be lost altogether. An illustration will assist in visualizing the loss of information that may result from the use of lower resolutions.

Information conveyed by Figure 1, an extremely detailed photograph, would undoubtedly be lost if its digital surrogates were created using lower resolutions. Edge blurring would prevent a researcher from studying wheels, spokes, and hubs, and clothing detail would become invisible. The

Figure 1. Ford Auto Dealers Auto Parade, 1912 (neg. 0.3299). From the collections of the Henry Ford Museum & Greenfield Village.

wide range of tonal contrast across very limited spaces would also be obscured, and the overall effect would be a smoothing of shadows and features.

Compression is a technique used to reduce the size of a digital file. This is accomplished in a number of ways, including mathematical transformations and reduction of precision by the elimination of "noninformational" data or noise in the data set (Brown & Shepherd, 1995). Reduction of precision is the most commonly applied compression technique; the effect that it has on the quality of the resulting digital image makes compressed files very attractive to institutions that are concerned with potential unauthorized use of images. Some museums believe that, as in the case of employing low resolution, reducing the quality of the digital image file makes it unattractive to would-be electronic privateers. Reducing the size of a file means that the quality of the resulting image is reduced and that the amount of time that it takes to transmit the data from the file over a communications link is decreased. Using compressed files for Internet or intranet transfers therefore is doubly attractive, but there is a potential loss of image integrity that may prove significant depending on the use of the files. The risks are more obvious when compression algorithms are examined in greater detail.

The class of compressed formats termed "lossless" is based on data transformation algorithms (for a discussion of wavelet and fractal compression, see Puglia, 1998). In these formulas, the original scanned pixel values are transformed into other values, most often using either run-length encoding (i.e., Sunraster, TARGA, and TIFF format types 2 and 32773), LZW encoding (i.e., GIF LZW and TIFF scheme 5), or discrete cosine transforms, also known as DCT (i.e., JPEG DCT and MPEG DCT). One-dimensional differencing, a method employed to produce JPEG predictive implementation, a true lossless JPEG format, is not discussed here. In run-length encoding, repetitive sets of identical data values in the original data are replaced by codes, each made up of a single data value and a length value. The collection of codes and their values must be stored in the file as the "codebook." The resulting reduction in file size depends on the number of repetitive data sets in the original. The use of the compressed data is contingent on the ability of the user software application to use the codebook to decode the format.

Lempel-Ziv and Welch developed an alternative method (LZW) of encoding data in 1985 that also uses pattern recognition but allows the decoder to build the codebook as it processes the data stream. The resulting file is smaller than those created with run-length encoding. In formats that use DCT to compress files, a mathematical operation is applied to blocks of original data. The transformed block is represented by fewer bits in the digital file than the original block. Run-length and LZW encoding result in no loss of data; DCT does in fact result in some data loss due to round-off errors, but the overall effect on the quality of the resulting image is inconsequential. DCT tends to yield higher compression ratios; nevertheless, average ratios of original data file size to compressed file size tend to fall in a range of 2:1 to 9:1 (Brown & Shepherd, 1995, p. 190). TIFF and lossless or near-lossless JPEG formats are extremely attractive for the purposes of maintaining data integrity, but their application does not result in files that are as small as those created employing other "lossy" techniques.

Reducing the precision of data means eliminating information in the original file that is not necessary for the purpose at hand. The electronic scanning of a photograph, for example, may produce sets of greyscale values that, while different, are so close to one another in tonality that the human eye may not be able to distinguish a difference. Recording data at this level of precision is probably not necessary for the creation of an acceptable digital surrogate. Achieving more aggressive compression ratios, in the area of 20:1 and higher, requires the establishment of less stringent definitions of noise and results in more notable erosion of information content. Commonly used implementations of JPEG (there are twenty-nine total) use reduction of precision combined with other data encoding techniques to achieve compression ratios up to 100:1. In these

implementations, DCT is used to transform the original data block information. A process called quantization is then applied to the transformed information; in this step, the transformed data are encoded as the result of rounding an amount produced by dividing the original value by some quantizing factor. Manipulating the quantizing factor effectively changes the amount of space that is needed to store the results of the quantizing process. Establishing the quantizing factor sets a threshold that divides data which are considered useful (and therefore are more faithfully retained) from data that are considered noise (and therefore are discarded). However, one person's noise may be another person's meaningful information.

Compression ratios of 32:1 can still produce images that are useful for some applications. The nature of the image source should be evaluated, however, to determine if highly compressed derivative files represent the original accurately enough to be acceptable for use. The photographs in Figures 2 and 3 are examples of source images that may not be acceptably represented by highly compressed digital image files.

Figure 2 depicts an open ledger book with written entries. The contrast between the handwriting and background is high, but examination of the blank areas of the ledger pages reveals that there is a fairly uniform layer of smudgy fingerprints that covers the page surface. Quantization of the original data from a scan of this image will undoubtedly result in the loss of this information. Given the nature of the artifact, this would have a definite effect on any interpretations based on a study of images generated from a compressed digital surrogate.

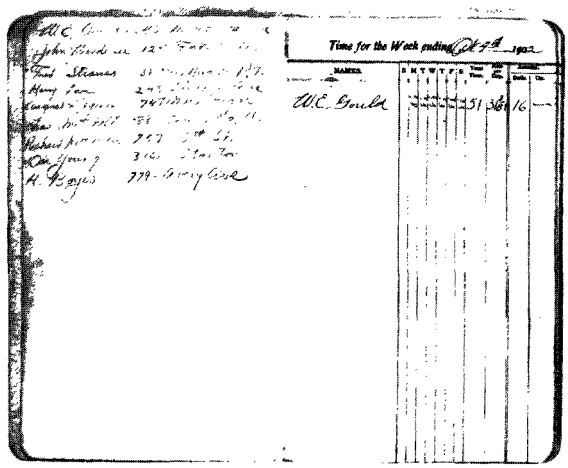


Figure 2. Payroll ledger entry for first Ford Motor Company employee, 1902 (neg. D.675). From the collections of the Henry Ford Museum & Greenfield Village.

In Figure 3, the pocking of the glazed finish creates a uniform stippled pattern across the surface of the jug. On the jug, there is incised ornamentation in the ship that emphasizes the stippled effect. Compression of original image data scanned from this photograph would result in a reduction of tonal contrast across the jug and subsequent loss of fine detail in the resulting image surrogate.

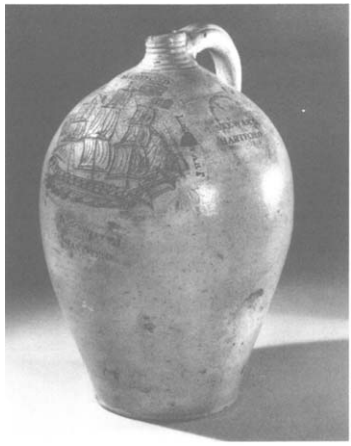


Figure 3. Salt-glazed stoneware jug, 1820–1830 (accession 57.67.8; neg. B.16963). From the collections of the Henry Ford Museum & Greenfield Village.

Purposeful Modification of Digital Image Information

Museums and archives that were early adopters of digital image technology often discover that the electronic representations created in the first years of the digital revolution are less than satisfactory when compared to those produced with current technology. In the case of the Henry Ford Museum, an early version of the automated collections management system was designed to work in tandem with laser disk readers. Laser disk images were created from 90,000 photographs from video documentation. Now transferred to PhotoCD, even the good images (degraded by processing three steps removed from the original) are difficult to use without some manipulation. Color modification is the most obvious intervention that is applied to older digital image files. Commercial image manipulation software packages provide a variety of other techniques to modify file information. At times there are side effects caused by the image processing operations that are used to enhance a problematic image in the form of modifications of image information that does not directly relate to the condition being corrected. For example, noise suppression using a method called Gaussian smoothing often results in the blurring of edges on shapes within the image (Davies, 1997, p. 44).

Furthermore, image enhancement operations that result in changes of brightness or contrast, noise reduction, and the sharpening of edges may employ filtering and thresholding algorithms that cause edges to shift in position and image shapes to become distorted. Curves and circles are particularly susceptible to shift. In images that contain both straight and curved edges, shapes may appear to move in relationship to one another after noise suppression filtering is applied (p. 59).

Attempting to correct noise or modify contrast in an image based on data scanned from Figure 4, a photograph of the Ford Rouge Plant, could result in distortion of shapes and subtle changes in the perspective of the image. If this occurred, the digital representation would present a false picture of the location of the camera, the size of components, and their spatial relationship to each other.

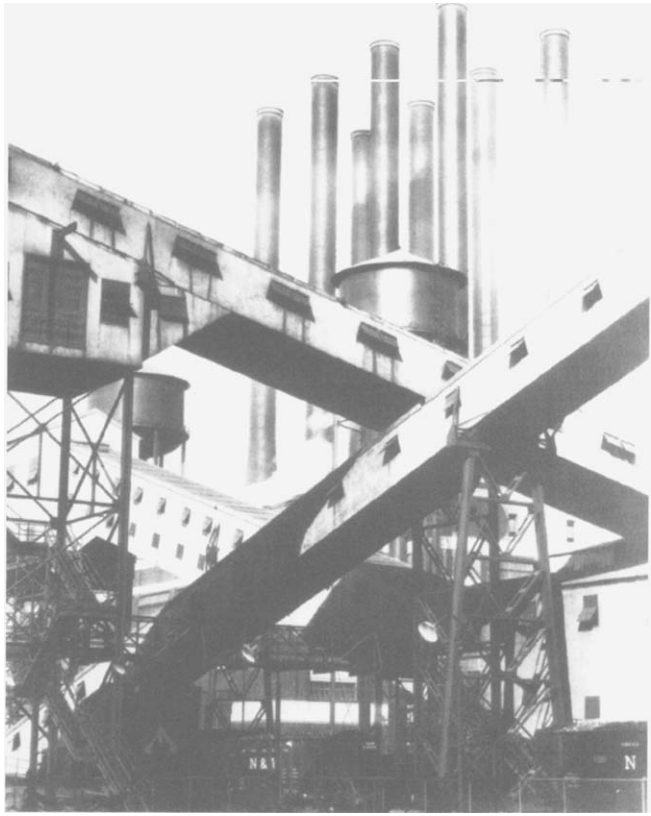


Figure 4. Conveyor System and Power Plant Stacks, Ford Motor Company Rouge Plant, 1927. Photograph by Charles Sheeler (neg. B.189.6577). From the collections of the Henry Ford Museum & Greenfield Village.

Handwritten letters, old photos, and artifacts are not the only collections items documented digitally in museums and archives. Figure 5, an image of a Model A parts drawing, was printed from an image in a collection of large-format microfilms, the only existing copies of this and other drawings. The original drawings and production copies were destroyed; the microfilm is the only remaining resource for specifications used in the reproduction of authentic parts. Any edge shifts or spatial distortions caused by manipulations of the digitized versions of these drawings could lead to disastrous misinterpretation of the images by the parts manufacturers who use them.

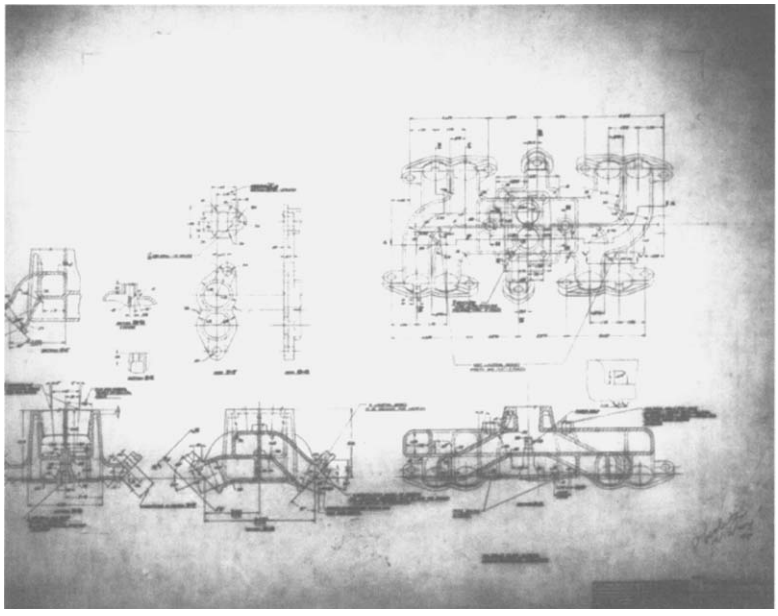


Figure 5. Model A Parts Drawing, 1944. From the collections of the Henry Ford Museum & Greenfield Village.

Most purposeful image data modifications cause the existing image characteristics to change or disappear altogether. A proactive security measure taken by some institutions to protect ownership is the addition of digitized credit line information that is either superimposed on or appended to the original image data. The thumbnail images displayed in Just In Time Images, the photo reproduction page on the Henry Ford Museum & Greenfield Village Web site, are altered in this fashion (http://www.hfmgv.org/jit/still/index.htm). The overlaid information obviously takes the place of the image data that formerly occupied that space. In the event that credit information is appended to an existing

image, the addition of pixels to the existing file results in the deletion of pixels from some other location in the image. The usual result is a cropped image. Electronic cropping may also be a purposeful action, prompted by display size limitations or aesthetic considerations. Regardless of reason, the elimination of digital information results in changed orientation of the image elements to the boundaries of the image.

The measuring rule normally included in documentation photos has been cropped out of this image of a single artifact (Figure 6). As a result, there is no referential information to provide a sense of the scale of the subject. Cropping can also remove spatial reference points that typically occur near the edges of images such as horizon lines.

Figure 6. Painted tin box, ca. 1825 (accession 29.737.2; neg. B.62832). From the collections of the Henry Ford Museum & Greenfield Village.

Failure to detect corruption of digital image information means that invalid, partial, or inappropriate information will be spread under the guise of authentic reliable information. It is important to re-emphasize that variation in data integrity among multiple surrogate versions of the same image is acceptable because of storage size, level of detail, and delivery speed requirements imposed by different uses.

The Role of Metadata in Maintaining Image Integrity

It can be argued that any information lost as the conscious or unconscious result of the processes described earlier could be restored intellectually if the digital image information is associated in some way with contextual metadata. Museum collections management systems usually display thumbnail images on the same screen as catalog information for the artifacts that are documented by those images. The text descriptions of the artifacts, if sufficiently rich and detailed, assist in the interpretation of the image information and vice versa. Image file technical metadata assists

in the evaluation of the digitized image based on the nature and limitations of its method of creation. Yet the two sets of complimentary information, metadata and digitized image, are usually stored in separate files. The catalog data exist in a series of rows linked across the tables of a relational database by virtue of unique identification numbers. The image data are stored in discrete files, one for each digital representation. The names of the files often bear no resemblance to the name or the accession ID of the original object. The connection between the sets is a one-way street leading from the text data to the digital image file. It is also fragile; the loss or modification of data in a single image file name field prevents the user of the catalog from viewing the image as well as preventing the user of the image from viewing associated catalog entries. If the connection is somehow broken and if the sets represent hundreds of thousands of artifacts or manuscripts, it will be almost impossible to properly relink all of the records and files. Standards related to the composition of contextual metadata aside, there is a serious need to consider the adoption of image data file formats that in some way automatically incorporate metadata in their structure. There are numerous informative discussions on the topic of metadata in museum and archival applications available both in print and on the World Wide Web (http://www.cimi.org; http://www.gi.getty.edu/index/warwick.html; http://www.acctbief.org/avenir/images.htm).

Ownership, Copyright, and Control of Digital Image Assets

In a world of increasingly complex legal issues, few pose more varied and vexing problems than those surrounding copyright and the ownership of images and image surrogates. Copyright laws were created to protect the rights of individuals to own the expressions of their ideas (Malaro, 1985, p. 113). Copyright is actually a suite of rights that may be conveyed, transferred, or retained, singly or in sets. Copyrights include (a) the right to reproduce the work, (b) the right to produce derivative works from the original, (c) the right to distribute copies for sale, (d) the right of performance, and (e) the right to display the work. Before 1978 in the United States, copyright existed only if the artist distributed the work with the copyright symbol; failure to do so was deemed a waiver of copyright. Copyrights to works acquired by a museum were assumed to transfer to the museum unless specific statements were made to the contrary. After the revision that took effect in January 1978, copyright was considered implicit in the act of creation and could only be waived by a statement to that effect. Museums can no longer assume that rights transfer automatically.

Until recently, expression of ideas implied the act of creating something with physical presence: a book, a painting, or a better mousetrap. Rights of authorship could not be enforced without recourse to referencing

something tangible or a tangible copy of a work. Digital representation is not easily categorized as having physical presence; there is no question that original work is involved, but marking or branding or seizing control of the "thing" that is created either as an original work or copy is conceptually difficult. As John Barlow (1996) describes the situation, under original copyright law "the bottle was protected but not the wine. Now the bottles are vanishing" (p. 11). Digital assets are the wine without the bottle. Controlling the use of digital image information representing items to which the museum clearly has copyright is difficult due to the accuracy with which duplicates can be made and the speed with which they can be disseminated (Bearman & Trant, 1997). Ambiguity regarding rights to photographic and digital reproductions of works in the public domain further complicates the process of enforcement and control (Akiyama, 1997). These reproduction rights, historically defended by museums and used to generate licensing income, have been threatened by a recent court decision that has implications for the control and use of digital reproductions. In this case, the Bridgeman Art Library, a British company that licenses transparencies of public domain art works that are owned by museums and collectors, brought suit against Corel Corporation, makers of a CD-ROM product containing digital reproductions of well-known paintings including 120 from the Bridgeman portfolio. Corel neither licensed nor requested permission from Bridgeman to use the works over which Bridgeman claims to have sole control. Bridgeman maintained that Corel had violated their copyrights; Corel countered by claiming that the museums and collectors could not assign to Bridgeman the rights that pertain to works in the public domain. The court ruled in favor of Corel, finding that substantially exact photographic reproductions of two-dimensional works of art are not copyrightable because they do not involve original work ("Copyright Case," 1999). The implications of this decision are serious. If it is upheld, museums will neither be able to exercise control over the use of image reproductions of public domain items in their collections nor to charge copyright fees for the use of such images, no matter the format.

Assuming that an institution's rights to the digital representation of an image are established, copyright enforcement can be accomplished in two ways. The institution can control access to the digital file permitting use only by those who are appropriately authorized. On the other hand, the museum or archive can provide unlimited access to digital visual resources that are marked with an indication of proper ownership. Suspect reproductions can then be examined and, if the mark is detected, the institution can proceed with steps to enforce their rights.

Most collections management software packages provide users with one or more security schemes. These are implemented by logon id and selectively allow each user to perform pre-defined sets of operations on

specific fields or files. Application system security can be an effective way to prevent the modification of text-based contextual information. Digital envelopes can also be used to protect text files that contain metadata relating to digital image files. A digital envelope uses encryption to permit access to file content on a selective basis. The text data are encrypted using a key and then the key itself is encrypted using another key. The user must decode the key data before it can be applied to the content in a second decoding step; double keys protect the content from both casual theft and from most true data pirates.

Digital image information is stored in discrete files separate from a text-finding aid or catalog information. The image files are accessible to proprietary system users and to everyone else with image server access as well. The most recent implementations of Microsoft Windows provide image display capabilities as default readers that respond to Open commands; the user can invoke them by selecting the image file name from any storage device. These files can be modified using any commercial image processing software. While it is possible to store image file information in a digital envelope (Acken, 1998), this technique requires that all potential legitimate users be identified and equipped with appropriate keys. It is often undesirable (as in the case of images used on a Web site) to prevent casual viewers from seeing an image. In this case, marking the image files and monitoring their use is an effective way to protect content and enforce copyrights if necessary.

Digital watermarking is the process of inserting marks or labels into digital content in such a way that they are unobtrusive yet inseparable from the source data (Yeung, 1998, p. 32). This article has already referenced the use of visible credit lines superimposed over the source image. This technique visibly alters the content of the surrogate image, displacing potentially meaningful data. Most digital watermarks are transparent. There is no degradation of visible content caused by the watermark, but the watermark is detectable using special software processes. A good analogy can be drawn from photocopying. At the Henry Ford Museum Research Center, copies of photographs or documents from the collections are made on a photocopying machine using paper that is pre-stamped with a rights and use warning in red ink. The red message displaces meaningful data from the source document. If watermarked bond paper without the stamp was used for the copies, no meaningful source data would be displaced but the watermark could still be viewed under certain circumstances, as on a light table.

The analogy breaks down when a reproduction of the photocopy on bond paper is made. The watermark will not appear as part of the photocopied information, although the overall quality of the content will degrade as copies are made from other copies. In digital watermarking, the file contents can be duplicated an infinite number of times with no

degradation of quality, and theoretically the watermark will appear the same in every copy.

Watermarking technology is opportunistic, relying on the fact that in any digitized image file (including compressed files) there are some bits that carry less significant information than other bits. In invisible watermarking, modification of the data in these bits causes minimal visible change in the image when displayed or printed (Memon & Wong, 1998). The modifications are data substitutions that collectively make up the watermark. The degree to which the discernible image content is affected depends on the nature of the image (if it contains large areas of solid intense color for example) and the nature of the watermarking algorithm (Wayner, 1997). Visible watermarks affect the image, usually by adding a transparent logo or visual message to the displayed data. In both types of watermarking, modified data are located in different places or "holes" in the image file and can be extracted and assembled to convey meaningful information.

There are ways to remove watermarks but benchmark standards for robustly resistant watermarks are being developed (Mintzer et al., 1998). Robust watermarks are those which can be recovered in spite of intentional or unintentional modification of the image file. They must be able to survive a variety of processes including filtering, cropping, scaling, and compression. This type of watermark is useful for establishing ownership of an image or for detecting unauthorized copies. There is another form, fragile watermarking, that relies on the ability of the mark to break easily if the image is altered. Fragile watermarks are designed as tools for identifying compromised data; they can even shed light on the nature of the alteration. If the ruling in the *Bridgeman vs. Corel* case is not struck down, fragile watermarking may be the only way to ensure that uncontrolled use of digital files does not result in a proliferation of inaccurate and unauthentic images on the World Wide Web.

Developing a Digital Image Risk Management Plan

Risk management plans should be developed based on the unique nature of each institution's digital image holdings and the audiences that access them. It is useful to recall that the purpose of risk assessment is to develop acceptable accommodations of failure. Perfection is neither obtainable nor necessary. Few, if any, institutions have the resources to create, store, and use images that are perfect electronic replicas of the originals by current standards. It is enough to be aware of the compromises that are made and of the impact that they may have now and in the future. It is also important to be informed and open to change.

There are a number of development efforts underway that could have a major impact on the manner in which museums and archives use and distribute digital image information. One example of an exciting emerging

technology is the FlashPix image file format, developed by a consortium of high-tech companies including Eastman Kodak, Hewlett-Packard, Live Picture, Inc., and Microsoft (Donovan, 1998). FlashPix addresses a number of problems. It allows the storage of original digital input plus a number of lower resolution copies in the same file. Each resolution is broken into smaller segments called tiles that can be read individually or in groups. FlashPix also allows text metadata to be stored in the same file as the image data, solving the problem of developing standardized headers or maintaining links between image files and contextual data stored elsewhere. Although not currently supported in browser software, this format could greatly simplify the digital image risk management process.

Digital watermarking technologies are also changing rapidly. Commercial applications are concentrating on rights enforcement and signature authentication applications, but there is a growing interest in using the "holes" in digital image files for the storage of metadata. One author suggests that embedded hyperlinks could direct viewers to related Web sites and that embedded indexing data could be used to pre-select images for viewing (Acken, 1998 p. 77).

There is an obvious need for ongoing research, evaluation, and planning if museums and archives are committed to protecting their digital image assets. A number of potential threats to the integrity of digital image information have been identified here. Changes in the integrity of digital image information can be caused by the manner in which the source data are acquired and recorded and by modifications made to the image data file. Alterations made to contextual data can limit valid interpretation of the associated surrogate image. The destruction of the mechanisms that link contextual data to the appropriate digital image has the same effect as deleting contextual information. Loss of control over digital assets can be the result of failure or inability to establish and publicize copyright. Even if copyright is established and enforceable, failure to enforce rights has the same effect as having no rights at all. Finally, failure to detect corruption of digital information means that invalid, partial, or inappropriate information will be spread under the guise of authentic reliable information. Some institutions are already proactively applying security measures to digital image collections. As noted here, security measures can have a negative impact on the integrity of the files that they are designed to protect. Systematic consideration of risk factors can inform the creation of procedures and application of security that works to guarantee the reliability and accuracy of digital image assets.

REFERENCES

Acken, J. M. (1998). How watermarking adds value to digital content. *Communications of the ACM, 41*(7), 74-77.

Akiyama, K. A. (1997). Rights and responsibilities in the digital age. *Visual Resources, 12*(3-4), 261-268.

Barlow, J. P. (1996). Selling wine without bottles: The economy of mind on the global net. In P. Ludlow (Ed.), *High noon on the electronic frontier* (pp.1-8). Cambridge, MA: MIT Press.

Bearman, D., & Trant, J. (1997). Museums and intellectual property: Rethinking rights management for a digital world. *Visual Resources, 12*(3-4), 269-280.

Besser, H., & Trant, J. (1995). *Introduction to imaging: Issues in constructing an image database.* Santa Monica, CA: Getty Art History Information Program.

Bridgeman copyright case update. (1999). *Aviso* (April), 1.

Brown, C. W., & Shepherd, B. J. (1995). *Graphics file formats: Reference and guide.* Greenwich, CT: Manning Publications.

Conway, P. (1996). *Conversion of microfilm to digital imagery: A demonstration project: Performance report on the production conversion phase of Project Open Book.* New Haven, CT: Yale University Library.

Copyright case challenges long-held museum assumption. (1999). *Aviso* (February), 1.

Davies, E. R. (1997). *Machine vision: Theory, algorithms, practicalities.* San Diego, CA: Academic Press.

Donovan, K. (1998). The promise of the FlashPix image file format. *RLG Diginews, 2*(2). Retrieved October 1, 1999 from the World Wide Web: http://www.rlg.org/preserv/diginews/diginews22.html#FlashPix.

Frey, F. (1997). Digital imaging for photographic collections: Foundations for technical standards. *RLG Diginews, 1*(3). Retrieved October 1, 1999 from the World Wide Web: http://www.rlg.org/preserv/diginews/diginews3.html#com.

Kenney, A. R. (1997). The Cornell Digital to Microfilm Conversion Project: Final report to NEH. *RLG Diginews, 1*(2). Retrieved October 1, 1999 from the World Wide Web: http://www.rlg.org/preserv/diginews/diginews2.html#com.

Malaro, M. C. (1985). *A legal primer on managing museum collections.* Washington, DC: Smithsonian Institution Press.

Memon, N., & Wong, P. W. (1998). Protecting digital media content. *Communications of the ACM, 41*(7), 35-43.

Mintzer, F.; Braudaway, G. W.; & Bell, A. E. (1998). Opportunities for watermarking standards. *Communications of the ACM, 41*(7), 56-64.

Puglia, S. (1998). Fractal and wavelet compression. *RLG Diginews, 2*(3). Retrieved October 1, 1999 from the World Wide Web: http://www.rlg.org/preserv/diginews/diginews23.html#technical2.

Reilly, J. M. (1995). Technical choices in digital imaging: The Technical Images Test Project in review. In P. McClung (Ed.), *RLG Digital Image Access Project* (Proceedings from an RLG symposium held March 31 and April 1, 1995, Palo Alto, CA) (pp. 85-93). Mountain View, CA: Research Libraries Group.

Wayner, P. (1997). *Digital copyright protection.* Boston, MA: AP Professional.

Yeung, M. M. (1998). Digital watermarking. *Communications of the ACM, 41*(7), 30-33.

Additional References

CNRI Registry. (1998). *Handle systems overview.* Retrieved October 1, 1999 from the World Wide Web: http://www.handle.net/overviews/hs-version4.html.

Consortium for the Computer Interchange of Museum Information. (1999). *Home page.* Retrieved October 1, 1999 from the World Wide Web: http://www.cimi.org.

Getty Information Institute. (1997). *Metadata standards.* Retrieved October 12, 1999 from the World Wide Web: http://www.getty.edu/gri/standard.

Image Quality Working Group of ArchivesCom. (1997). *Technical recommendations for digital imaging projects.* Retrieved October 1, 1999 from the World Wide Web: http://www.columbia.edu/acis/dl/imagespec.html.

Library of Congress Preservation Office. (1998). *Manuscript digitization demonstration project final report.* Retrieved October 12, 1999 from the World Wide Web: http://memory.loc.gov/ammem/pictel/index.html.

National Digital Library Program, Library of Congress. (1998). *Digital formats for content reproductions.* Washington, DC: C. Fleischhauer. Retrieved October 12, 1999 from the World Wide Web: http://memory.loc.gov/ammem/formats.html.

Sandore, B. (1997). Images and their descriptive metadata. In *Proceedings of the Conference on the Future of Communication Formats* (Held in Ottawa, Canada, October 5-10, 1996, sponsored by the Banque Internationale des Etats Francophones and the National Library of Canada) (pp. 121-133). Ottawa, Ontario, Canada: Banque Internationale des Etats Francophones.

Simmons, C. (1998). *Risk management.* Retrieved October 1, 1999 from the World Wide Web: http//www.airtime.co.uk/users/wysywig/risk_1.htm.

Getting the Picture: Observations from the Library of Congress on Providing Online Access to Pictorial Images*

CAROLINE R. ARMS

ABSTRACT

OVER THE LAST FEW YEARS, THE LIBRARY OF CONGRESS (LC) has increasingly created digital reproductions of visual materials to enhance access to its resources. Digitization is now a mainstream activity in the Prints and Photographs Division (P & P) and the Geography and Maps Division (G & M). Both divisions work closely with the National Digital Library Program to make their incomparable resources accessible over the Internet to the general public through the American Memory Web site (http://memory.loc.gov/). They also use the digital images to serve their more traditional clientele in the reading rooms. Retrieval from a collection of digital images offers special opportunities to apply new technological advances, as illustrated elsewhere in this issue. However, retrieval often takes place in broader contexts. The Print and Photographs Division seeks to enhance access to its international pictorial holdings, whether digitized or not. Within American Memory, the focus is on retrieval by the nonspecialist from a body of materials related to the history and culture of the United States, materials heterogeneous in both original and digital form. A yet broader context is retrieval from the comprehensive collections of the entire Library of Congress. Beyond enabling retrieval, LC is concerned with facilitating use of the materials retrieved, consistent with any associated rights. This article describes selected aspects of LC's practical experience and current practices from digital capture through interactions with users, with an emphasis on the integration of access to pictorial images online with other services and activities at LC.

Caroline R. Arms, Library of Congress, 101 Independence Avenue S. E., Washington, DC 20540-9300

LIBRARY TRENDS, Vol. 48, No. 2, Fall 1999, pp. 379-409

Considering the Challenges of Access and Retrieval

Choices made at the Library of Congress in relation to access, retrieval, and use of its pictorial materials reflect many requirements and desires:

- to serve different audiences from expert researchers to the K-12 and higher educational communities and the lifelong learner;
- to support a variety of uses as appropriate, including citation (in print and online as active hyperlinks), study and comparison, convenient reproduction for classroom or personal use, and high-quality reproduction for publication;
- to facilitate access to digital pictorial resources in conjunction with access to related materials in all forms;
- to find a balance between demand by users for ever more detailed description for resources currently accessible and for access to more of its collections;
- to allow digitization to serve a future role in long-term preservation of materials originally created in many forms;
- to find practical solutions in the absence of well-established standards and contribute to the informed development of standards where necessary; and
- to build systems that can be deployed today with large quantities of images and to enhance services incrementally taking into account economic and organizational realities.

The pictorial collections of the Library of Congress present enormous challenges for both physical and intellectual access. The Prints and Photographs Division (P & P) holds over 13 million images, including photographs (published and unpublished), cartoons, posters, documentary and architectural drawings, and ephemera, such as baseball cards. Most of the images are photographs related to the United States, many acquired in large collections. Photographic prints may have been captioned (usually by writing on the physical artifact) or organized by the photographer or by the institution or individual from whom a collection was acquired. Many images, however, are held only as negatives, which pose special problems for identification, housing, and service and are not always accompanied by individual captions.

Cataloging pictorial items is labor-intensive. P & P estimates that it takes an hour to produce a brief record for an average item for inclusion in the Library of Congress' main catalog. This allows time to handle the item appropriately, note identification numbers, record basic information about the creator, date, physical artifact and reproduction rights, devise a descriptive caption where necessary, assign a few subject headings, and proofread. One hour stretches to three or four if an attempt is made to verify the information accompanying the piece, to describe what a picture

is about rather than simply what it is of (consider a political cartoon or a photograph of a notable event), or to provide added contextual information, such as where a picture was subsequently published or biographical notes on the subject of a portrait. This degree of effort can only be justified for a small portion of the P & P holdings—e.g., for fine prints and posters.

Before machine-readable cataloging, access to the Prints and Photographs Division resources was primarily through a card catalog with entries for groups of material and through "browsing" files organized in thematic hierarchies but with no separate item-level description or control. Because of the size of the collections, many items are still only accessible this way. Since 1989, a priority for the Library of Congress has been reducing the backlog of items waiting to be included in public access systems; efforts in P & P since then have focused on physical organization and cataloging for materials that were previously unprocessed or in high demand.

Visual approaches to browsing pictures offer an alternative to detailed cataloging of individual items. Once a picture is retrieved, however, most users need information about the picture in order to cite it, confirm its applicability as evidence or illustration, or determine whether permission is needed to reproduce it. At the very least, the user wants to be able to find a particular picture again (preferably directly rather than by browsing), request a reproduction, or seek permission to reproduce it. In the Prints and Photographs Division reading room, a variety of clues and experts are available to allow identification. Item-level control may be deferred until demand for the particular item is demonstrated. When a user requests a reproduction of an item that has not previously been cataloged, it is assigned a reproduction number and an item-level record is created. Digitization accelerates the need to apply identifiers to the physical items, both for tracking during conversion and quality review and to relate the digital copy to its physical source.

The Prints and Photographs Division regards its digital reproductions, even the high-resolution images being prepared under the current conversion contracts, as surrogates. Based on his experience as chief of the Prints and Photographs Division, Ostrow (1998) reviewed the nature of large historical pictorial collections and how they are traditionally used in reading rooms. He emphasizes the constructive role digital surrogates can play for researchers and in cutting down the need to handle fragile originals. He also discussed shortcomings of digital images as replacements when used for historical documentation. The Library of Congress has not yet used digital reproduction to replace physical originals, even when the originals will soon be unusable. To preserve the information for the longer term, brittle books are currently microfilmed and deteriorating negatives are replaced by high quality photographic copies.

For many uses and users, however, the digital surrogates suffice. Digitization has furthered the objectives of the staff in the Prints and Photographs Division to serve patrons in the reading room better and of the National Digital Library Program (NDLP) to make resources available to a much broader public beyond LC's walls. The remote audiences, however, present new expectations and a wider range of tasks for which pictures are needed; they also lack access to expert assistance. The challenges of serving many audiences and supporting retrieval in many contexts will continue. LC's current technical architecture is based on a modular framework that allows different interfaces to take advantage of the same catalog records and the same digital content. The same digitized picture can be accessed through LC's comprehensive catalog, through American Memory, or through a catalog that is tailored to pictorial resources.

Looking Back

Released for public access over the Internet in early 1998, the Prints and Photographs Online Catalog (PPOC) (http://lcweb.loc.gov/rr/print/catalog.html) is the most recent interface to an increasingly comprehensive catalog to the division's holdings. Where available, records in PPOC are accompanied by digital images. This catalog builds on work which started in 1982 when the division began reproducing selected collections electronically (initially on videodisc) and cataloging the images for LC's Optical Disk Pilot Program, described by Elisabeth Betz Parker (1985). In December 1993, a dedicated workstation with an array of videodisc players and a separate monitor for displaying images was introduced as a public service in the reading room and dubbed the "One-Box." The "One" in One-Box represented the goal of the Prints and Photographs Division to develop a reference gateway that could provide access to all their holdings. In 1996, a digital version (known initially as the Digital One-Box and taking advantage of the capabilities of the World Wide Web) was introduced in the reading room. The Digital One-Box became the Prints and Photographs Online Catalog and was released on the Internet after incremental improvements based on experience with users. By December 1998, PPOC provided access to twenty-five collections covering over 5 million physical items. In the P & P reading room, PPOC provides access to 355,000 digitized images. Over 60 percent of these images are accessible through American Memory; others are out of scope, or access is restricted because of copyright or other reasons, such as privacy. American Memory and PPOC share digital image files and catalog records for the overlapping content and rely on much of the same program code.

The American Memory pilot project in the early 1990s explored the use of digital images on CD-ROM. The Prints and Photographs Division participated actively in the pilot and, in June 1994, the first release of

American Memory on the World Wide Web comprised three collections of photographs. By December 1998, thirteen collections from P & P, representing 220,000 original items, had been released on American Memory.

Two very large collections are being digitized and released for public access in phases. The first, known as Built in America (http://memory.loc.gov/ammem/hhhtml/hhhome.html), comprises photographs, architectural drawings, and "data" pages of typed textual documentation from the Historic American Buildings Survey and Historic American Engineering Record (HABS/HAER). As of March 1998, HABS/HAER documented 35,000 sites and structures through 363,000 negatives and paper artifacts. The other large collection is entitled America from the Great Depression to World War II (http://memory.loc.gov/ammem/fsowhome.html). It contains approximately 165,000 negatives and transparencies from the Farm Security Administration and the Office of War Information (FSA-OWI).

Today's Snapshot

The National Digital Library Program was established in 1995 as a five-year program. LC management is currently considering how best to build on NDLP's achievements and incorporate digital content more extensively into its collections. Figure 1 is based on a diagram developed to describe the component activities and related systems that provide the infrastructure for producing or incorporating digital content into LC's collections and providing coherent access to those resources. The diagram is based on the experience of staff in the NDLP, in the custodial divisions whose content NDLP has helped digitize and provide access to, and in LC's central technology service organization, which has a small group of programmers building and maintaining the computer applications that support both American Memory and the Prints and Photographs Online Catalog. The framework in this diagram will be used to organize the observations and experiences described in this article.

Making Digital Reproductions

The Prints and Photographs Division has chosen to use expert contractors to prepare the digital reproductions of pictorial materials. The use of contractors allows the Library of Congress to take advantage of special equipment without the need to build in-house facilities to handle a wide variety of physical formats. Contractors are also better able to keep up with the latest technological improvements in hardware and develop specialized software that applies the latest techniques for capturing and processing large quantities of images, since the investment required can be allocated across many projects and customers. In early 1998, a multiyear contract was awarded for the generation of digital images for pictorial materials after evaluating responses to a request for proposals (Library of

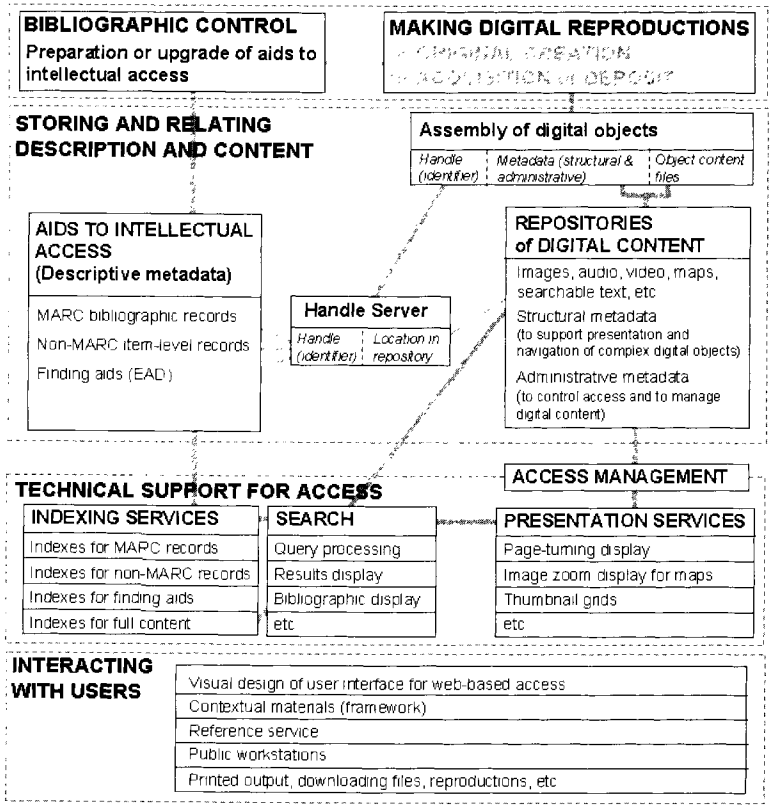


Figure 1. Infrastructure for Managing Digital Collections and Providing Access.

Congress, 1997, RFP97-9). The contract, awarded to JJT, Inc., covers a variety of original formats, including transmitted-light items (e.g., negatives and transparencies) and reflected-light items (e.g., photographic prints and baseball cards) but excluding oversize items such as architectural drawings. The RFP provides an excellent description of LC's objectives and the criteria considered important in digitizing pictorial materials. Some details, however, were modified during the contract startup and as production began. In line with the requirements to perform capture at LC, the contractor has established a scanning facility in a small room in the P & P division. On-site scanning allows items that cannot leave LC to be scanned directly rather than via photographic intermediates, and reduces manpower needs for shipping and tracking large quantities of material. Although LC used photographic intermediates for early projects, they inevitably introduce some degradation in quality as demonstrated in the RLG Technical Images Test Project (Reilly, 1995).

However much effort is devoted to describing specifications in a proposal or contract, LC has learned that other factors are important to the success of digitization projects. A cooperative working relationship with frequent communication is invaluable, as is the development of mutual trust. Early in this project, the contractor demonstrated a commitment to careful handling that allayed concerns of conservators. As indicated in a recent report from the Image Permanence Institute (Frey, 1998), operator expertise and visual sophistication are needed for successful digitization in an archival environment. After careful and productive experimentation with test batches of materials, LC is able to rely on the scanning contractor's technical judgment on many matters. The staff from the Prints and Photographs Division and the contractor's team have shared objectives, with ambitious goals for quality balanced by a need for productivity.

Quality of Digital Reproductions

The current Prints and Photographs Division practice is to scan most pictorial items at spatial resolutions of 3,000, 4,000, or 5,000 pixels on the long side. The choice depends on various factors, including the size and quality of the source and the visual content or intent of the work. For baseball cards (http://memory.loc.gov/ammem/bbhtml/bbhome.html), which are small, capture was at 3,000 pixels. The negatives in the HABS/HAER collection are being scanned at 5,000 pixels since they are intended to document architectural and engineering details and may be used in the future to support reconstruction or restoration. Scans from copy negatives are usually at 4,000 pixels since the quality of the copies does not warrant higher resolution. Anne Kenney of Cornell University and Lou Sharpe of Picture Elements, Inc. have used a conceptual structure for suggesting levels of resolution for capture of illustrations in nineteenth-century printed books in a study for LC's Preservation Directorate (Kenny et al., 1999). They consider the increasing resolutions needed to capture the essence of a picture for screen display, the detail of its visual content (such as a wisp of hair in a portrait), or the structure of the original artifact (for example, to distinguish different types of engraving). Although P & P does not use these categories explicitly, the aim is usually to capture the visual details rather than to reveal the artifactual structure. The equipment used by the contractor captures color images at 36 bits/pixel; any processing is at 48 bits/pixel to ensure that transformations do not introduce artifacts or lose detail. Due to limitations of current technology and standard formats, the tonal resolution is reduced to 24 bits/pixel for delivery to LC as an uncompressed TIFF image. Black and white photographs are captured at 12 bits/pixel grayscale, processed at 16 bits/pixel, and stored at 8 bits/pixel.

The file size for a color image at 5,000 pixels on the long side is around 50 Mb for the archival uncompressed TIFF. The archival grayscale HABS/

HAER images approach 20 Mb. Smaller derivative versions are created for convenient access, retrieval, and use, although the Prints and Photographs Division usually makes the archival versions of images not subject to copyright protection available for downloading. Current practice (which differs a little from the RFP) is to create three smaller versions. A thumbnail, 150 pixels on the long side for almost all items and at 8 bits/pixel, is designed for inclusion in item-level bibliographic displays and grids of thumbnails. This size delivers adequate performance over the Internet and supports rapid visual browsing on screen. Care is taken with the quality of thumbnails since a poor thumbnail may keep a user from looking at a relevant image.

For general use, two service versions are currently created. For convenient display on any screen and rapid downloading over the Internet, a JPEG image at 640 pixels on the long side is generated with moderate compression (usually at a reduction of around 15 to 1 for color and 8 or 10 to 1 for grayscale images). For the user who wishes more detail or a better image for printing or other re-use, a JPEG with lighter compression is created at 1,024 pixels on the long side. When creating the derivatives or during quality review, the contractor selectively applies techniques to reduce the moiré patterns that may be generated when images with regular patterns (such as siding on a house) are reduced in size. The technique most commonly applied is to blur the image at full size, reduce to the desired size, and then sharpen the smaller image.

The tonal quality of images is limited by the quality of the initial scan, which depends on the equipment, its calibration, the judgment of the scanning operator in using its capabilities, and environmental characteristics, such as dust and lighting. The scanning stations designed and installed by the contractor at the Library of Congress incorporate prototype MARC II digital cameras built by Udo and Reimar Lenz of Munich, Germany. The contractor has developed software to control settings, including the camera's height above the glass scanning surface which can be lit from below for transmitted-light materials and above for reflected-light items. No glass is placed directly on top of materials. Curtains and air filters around each station provide control over lighting and dust. Target images that allow objective measurement of scanner performance on spatial resolution and grayscale representation are scanned with each batch of material and stored.

For prints, the objective is to reproduce the tonality of items as they exist. The monitors on scanning and review stations are calibrated carefully, although it is recognized that no monitor can reproduce all tonal qualities of all originals in all viewing conditions. For negatives and original positive transparencies, the objective is to create a positive image using photographic sensitivities that might be expected of a skilled darkroom technician. For the documentary HABS/HAER images, the instructions

are to balance tones and capture all the information, avoiding loss of detail in shadows or highlights. Histograms of grayscale tonal values are used as an aid but not to control the process. All manipulations performed on each item, whether during scanning, quality review, or when creating derivatives, are recorded in a database by the contractor.

Archival images from collections scanned recently have been downloaded and reproduced in newspapers. For the February/March 1999 issue of *Civilization* magazine, a publication with higher quality requirements, archival digital images of baseball cards and some photographs were used. Some Prints and Photographs Division staff members argue that the high quality images being created currently could potentially serve as preservation masters in the future. LC has found it expensive to make photographic copies of deteriorating negatives. In addition, findings by the Image Permanence Institute suggest that scans from second-generation photographic materials are noticeably inferior to those from originals (Frey, 1998). The ability of the digital images currently being created from negatives to serve all the purposes served by photographic copies will be evaluated, and the practice of making photographic copies may be discontinued.

Workflow and Productivity

High throughput for scanning requires attention to workflow and modification of systems to avoid bottlenecks. The contractor has achieved an average rate of 375-400 negatives scanned in an eleven-hour day on a single station through a number of noteworthy innovations. One innovation is to plan and test transformations (including cropping and re-alignment) and exact steps for creating derivatives on small images (approximately 1,000 pixels on the long side) and then apply them to the full-size images automatically. A batch of these small images is sent on magnetic tape each night to the company headquarters in Texas, where highly skilled staff review the day's work and plan the transformations, which are recorded for automatic retrieval by the contractor's staff at LC. Another innovation relates to the size of individual files. Initially, operators were waiting while the large image file from one scan was stored before the preview of the next item could be displayed. The contractor modified the computer system to allow these tasks to run in parallel.

A limiting factor on throughput is now the capacity of the dual Pentium II workstations, which will be upgraded by the contractor as faster processors become available and can be tested. Another bottleneck is network bandwidth for loading the archival image files to the Library of Congress' servers. Rather than degrade network and server performance for users, image files are loaded directly to LC's storage system from CD-ROMs created by the contractor. Workstation capacity and network bandwidth have also been limiting factors for the Geography and Map Division (G & M)

where large format color maps are scanned by LC staff at 300 dpi and 24 bits/pixel, sometimes creating individual files of over 300 Mb. After each image is reviewed and cropped, a compressed version is created using wavelet compression. This post-processing generates a heavy computing load. Using 166 MHz Pentium stations with 32 Mb of RAM, the production rate was roughly six maps per day. New dual Pentium-II processors with 500 Mb of RAM have generated a sevenfold increase in throughput. In G & M, catalogers can still keep pace with the scanning operation. In the Prints and Photographs Division, however, the scanning capacity far exceeds the capacity for item-level cataloging.

Preparing Aids to Intellectual Access

Describing pictorial materials accurately is time-consuming and expensive. Unlike a book, which usually has a title page on which basic information is recorded, an image does not describe itself. Words are needed to indicate the place or event represented in a photograph, its creator, the names of people portrayed, and when it was taken. Providing effective access to large collections poses a challenge; unfortunately, many of the solutions for access in physical archives do not transfer as readily to the online environment as individual descriptions for each item, particularly when the aim is to provide coherent access to resources of all types.

The Prints and Photographs Division has made extensive use of collection-level and group-level records for cataloging its holdings. If a group of related items is housed in a single container, a single catalog record describing the contents will allow the user in the reading room to request the container and browse through its contents visually. Larger collections have sometimes been described both by collection- and group-level catalog records that point to a paper-finding aid that offers a structured hierarchical listing of the collection. Group-level catalog records for components of a large collection allow integrated general access from within LC's main catalog or the Prints and Photographs Online Catalog and identify the physical location and any paper-finding aid or index that offers more specific access. The finding aid provides a convenient mechanism for exploring the entire collection, once it has been identified as relevant, through the logical organization selected for its physical storage or any alternative indexes provided. As mentioned earlier, many items, particularly those acquired before automation, are accessible only through "self-indexing" filing schemes. For example, P & P has files for each president with subcategories such as cartoons, homes and haunts, and family. Biographical files hold portraits of many individuals, famous and not so famous. An advantage of such an approach is that the labor of preparing individual records for each picture is avoided. A disadvantage is that access is constrained by a single logical arrangement unless copies are made.

The physical filing scheme might seem straightforward to reproduce in a digital environment for visual browsing. However, browsing through a physical folder full of assorted pictures at a table in the reading room is not the same as browsing online. With the physical item in hand, the user will naturally scan the picture, turn it over to look for a caption, and make inferences from the physical nature of the item. In an online environment, some physical clues (such as size) are obscured and skimming through full-size images is awkward. Moreover, users have different expectations; online they expect captions to be legible and presented consistently. Even if a physical item has no individual identification, the user can ask a librarian about it by pointing to it. In an online environment to support remote users, the explicit recording of minimal identification for an image is essential if any appropriate use or reference is to be made. The challenge is to find lightweight approaches to description and organization that support both convenient visual browsing and also search-based retrieval on whatever descriptive information is available.

To date, items digitized from the Prints and Photographs Division's collections have mainly been described in item-level records, mostly in the MARC format, but with widely different depths of detail in the description. The first use of group-level records, described more fully below, has been made for the HABS/HAER collections. The records describe the intellectual expression and the original form of the material and provide a link to the corresponding digital reproductions. Information about the digital files is not recorded in the MARC bibliographic records since these are considered surrogates for reference purposes rather than separate works. One pragmatic reason for using the item-level approach initially is that it was easy to adapt existing search and retrieval tools and use the same records for PPOC, for American Memory, and for the main LC catalog. Another is that staff were confident that they knew how to design and build a first system around item-level records. Recent projects, at the Library of Congress and elsewhere, have begun to show how finding aids and group description can be used to advantage for providing access to pictorial materials.

In an ideal world, browsing and searching would be supported by item-level descriptions of uniform quality based on perfect information about time, place, and circumstances of creation, using descriptive terms from controlled vocabularies and following common practices for assigning subject terms. Rather than fixed browsing frameworks, groupings could be derived dynamically from the descriptive terms. Computing specialists designing systems often start by assuming that this ideal is easily achievable; to Prints and Photographs Division staff, who are responsible for large archival collections of pictorial materials, it is clearly not. They are nevertheless leaders in promulgating standards of practice since the ideal characteristics serve as goals and some are more achievable than others.

The division is responsible for a manual that supplements the Anglo-American Cataloguing Rules with details appropriate for graphic materials (Betz, 1982), which is included in the *Cataloger's Desktop* CD-ROM (Library of Congress, 1996), and maintains the *Thesaurus for Graphic Materials* (Library of Congress, 1995). Betz (1982) indicates that an individual image often derives its importance from the collection of which it is part, and that full cataloging may not be feasible for all works. For each collection or project within P & P, plans specify the level of granularity (e.g., item-level or group-level) and detail to be employed for cataloging. Factors considered include the research value of the material, its uniqueness, demand (past and potential), practicality, available resources, and how well proposed approaches fit with current systems.

Balancing Quality and Quantity in Practice

Faced with considerable quantities of material to which general access was unavailable, even in the reading room, and the Library of Congress' wish to provide online access for the general public to its collections to the degree possible, the Prints and Photographs Division has found ways to balance the pressures for both full description to support precise retrieval and access to a greater proportion of the holdings.

P & P recognizes different degrees of cataloging quality. Full cataloging, with verification of all information, addition of names to LC's name authority file, preparation of descriptive summaries, and extensive application of subject terms has been used traditionally for collection- and group-level records and for item-level records for certain categories of material, such as fine prints and posters. To prepare a full catalog record for an item takes between three and four hours. Of the collections within American Memory, only the Selected Civil War Photographs (http://memory.loc.gov/ammem/cwphome.html) and the daguerreotypes in America's First Look into the Camera (http://memory.loc.gov/ammem/daghtml/daghome.html) have full cataloging. In contrast, when a user requests a reproduction of an uncataloged item, "minimal" cataloging is performed for the item. Names of people and places are checked against the name authority file in the Library of Congress catalog; prescribed forms are used if found, and conflicts are avoided, but the research necessary to support the addition of a new name to the authority file is rarely performed. A small number of terms are chosen from the Library of Congress *Thesaurus for Graphic Materials (LCTGM)* to characterize the format and genre of the work and provide topical access. Whenever possible, a geographic heading is chosen from Library of Congress Subject Headings (LCSH). For minimal cataloging, no additional research is performed and no attempt is made to describe what the picture is about as opposed to what it is of. Such records typically take two hours to create and review. Reproduction requests for roughly 3,000 new items are received each year;

copy negatives created before this practice started are gradually being cataloged at the minimal level.

For most photographic collections selected for digitization as a whole, the Prints and Photographs Division has cataloged items at a preliminary level using only information from the piece. Captions and dates provided by the photographer are recorded but not verified. Where no specific information is available, a brief descriptive title is devised (and indicated), and an approximate date or date-range is given. Subject terms are applied sparingly, although consistently within a collection. Names are not checked against the authority file. To speed workflow, P & P develops collection-specific automated procedures to complete fields for information that are common to all or many item-level records in a batch. Preliminary cataloging for the photographs of Theodor Horydczak (for the American Memory collection Washington As It Was) took between thirty and forty-five minutes per record.

Large documentary collections, such as the HABS/HAER drawings and photographs that record architectural sites and engineering structures and the FSA-OWI photographs, pose particular problems for access. In making these available online for the public, P & P has explored new approaches to speed cataloging. Collections of negatives are often received by LC in an organization that roughly collocates pictures taken at the same time. For example, about one-third of the FSA-OWI collection was held on strips of 35mm negatives and many of the larger negatives were filed in the batches in which they had first been developed. The agency had provided information on photographers and geographic locations in various forms of documentation given to LC with the collection. For some images, captions and other notes had been recorded on cards. The information was transcribed by Library of Congress staff and merged with boilerplate information to create skeletal records in MARC format. Geographical locations were transcribed in the uncontrolled form used on the cards. Conversion to a standard form was performed automatically without verification. For roughly 35 percent of the images, no details were available beyond a sequenced call number. However, the caption or cataloging for one photograph often applies to or illuminates shots taken before or after it. To give users the clues offered by captions for neighboring images, a feature was added to American Memory and the Prints and Photographs Online Catalog that allowed visual browsing as a grid of thumbnail images sorted by call number.

For the HABS/HAER materials, the challenge was less the shortage of information to support access but more its incompatibility with library practice and formats. HABS and HAER are ongoing programs of the National Park Service, which keeps detailed information on the materials from each architectural site in a relational database. Every few months, batches of material are passed to the Library of Congress to serve and

archive. For this project, the Prints and Photographs Division took advantage of the fact that PPOC and American Memory use a general-purpose search engine designed to search several resources simultaneously, not necessarily in the same format. Routines were developed to convert the structure and format of the records in the relational database to a "flat" file of descriptive records more easily searched in combination with traditional bibliographic records. Each "bibliographic" record describes a site or structure. A single identifier provides a logical link to the content of all related documentation that has been digitized; hence the record serves as a group-level record. The documentation for a site may include photographs (with captions listed on separate pages), drawings, photographs, and pages of textual documentation (known as "data pages"). Each form of original content is being digitized independently using procedures for capture and quality review appropriate to the content. Whether the related content is yet digitized or not, records describing the documentation available are included in the database to support access to the physical collection and requests for reproductions. After new batches of image files are prepared and loaded, they are automatically retrievable from the bibliographic display.

Since HABS and HAER are ongoing programs generating new surveys, new copies of the database are retrieved regularly from the National Park Service, transformed, and re-indexed. No attempt is made to add controlled subject headings (although, with encouragement and assistance from the Library of Congress, the Park Service may start to use a controlled set of terms in light of its experience with providing public access). Much information from the original database is treated as notes; topical access is therefore primarily by free text search on the entire record. Automated expansion of query terms to include variants, which is done by default for both the Prints and Photographs Online Catalog and American Memory, is of particular value here.

For any HABS/HAER site, the related documents are treated as groups by original type. Hence, a typical engineering structure has an associated group of black-and-white photographs (with captions listed on separate pages), a group of drawings, and a group of pages of text digitized as page images. For each group, a small file (which LC has nicknamed a "page-turning data set") supports navigation through the group. The file holds sequencing information, links to component image-files, and optional captions. This data set, which can be thought of as "structural metadata" for the complex object representing the group of images, is generated automatically from names of files in a directory. The approach and the naming conventions that support the automatic derivation of the structural metadata were originally developed for American Memory, where it was first used to allow users to "turn" through pages of a short document, such as a theater program, and has since been extended to support page-

turning through much longer works. From each page, links to images of higher resolution permit more detailed study or better printing. The HABS/HAER drawings and data pages are presented through the original page-turning interface. For the photographs, the group is presented as a grid (or a sequence of grids) of thumbnails with associated captions. The programming for generating the page-turning data sets, the page-turning interface, and the thumbnail grid displays was done by Library of Congress staff.

The successful use of group-level description supported by thumbnail grids in HABS/HAER may lead to its use in future projects. There may be further opportunities to save cataloging effort by building on existing machine-readable descriptive records. The Prints and Photographs Division takes care to distinguish the levels of cataloging employed in records. Records that support access but do not meet LC's usual quality standards will not be distributed to the bibliographic utilities and other institutions as the basis for copy cataloging.

Preserving and Presenting Context

One characteristic of the Prints and Photographs Online Catalog and American Memory that is not provided in most library catalog systems, which developed primarily as tools for organizing holdings of independent monographs, is the clear delineation of a collection. The importance of a collection may lie in its overall scope and relationships among its parts; preserving the archival integrity of a physical collection is a guiding principle for curators. Online systems for archival pictorial collections should allow the user to wander within the boundaries of a collection with easy access to collection-level information that provides an intellectual context for the entire body of material. The front matter of an archival finding aid serves this purpose. American Memory and PPOC provide this context through an introductory framework of HTML pages, which often also feature exhibit-like presentations. This framework also allows LC to describe how a collection has been cataloged or digitized, in part to help users search more effectively or understand why all items do not have high-resolution reproductions and in part to inform others involved in similar ventures.

American Memory and PPOC encourage searching across the entire resource as well as within a chosen collection. To support the easy limiting of retrieval to item-level records from a single collection, collection identifiers (mnemonic codes rather than formal collection titles) are used (recorded in a local field in MARC records). The same coded field supports the automatic export of all records for a collection from LC's main catalog. More than one collection code can be included in a record, allowing the same item to be part of more than one collection. This is needed, for example, because items in LC's collections of daguerreotypes or

panoramic photographs may also be part of collections that relate to provenance. In the online environment, virtual collections can be assembled when curatorial or reference staff believe it will be valuable. The Prints and Photographs Division has used this capability for a few small selections made especially for American Memory. Examples include Votes for Women (http://memory.loc.gov/ammem/vfwhtml/vfwhome.html) and Jackie Robinson and other Baseball Highlights (http://memory.loc.gov/ammem/jrhtml/jrhome.html). More extensive use of the capability is made by the Geography & Map Division and the Motion Picture, Broadcasting, and Recorded Sound Division. For example, early motion pictures (many made by Thomas Edison) are presented as a body of material and as thematic collections; a separate virtual collection includes motion pictures made by Edison with his sound recordings.

Forms and Formats for Access Aids

The Prints and Photographs Division chose to use the MARC format for cataloging its pictorial materials primarily for compatibility with LC's main catalog and to allow distribution of records to bibliographic utilities (Zinkham, 1995, p. 48). The vision of a single catalog for all Prints and Photographs Division resources, and a wish to maintain only one copy of any descriptive record were other factors supporting this choice. Almost all records created and maintained by P & P are held in the MARC format, although the records may be derived automatically from a database used for processing the collection, as for the FSA-OWI photographs. Over the past few years, division staff have become skilled in exporting batches of MARC records to a form in which global changes can be made more easily and for converting back to MARC format.

The HABS/HAER records (for which the National Park Service maintains the data) are not converted to the MARC format but to a simpler format easily indexed by the search engine used for American Memory and the Prints and Photographs Online Catalog. This format identifies fields within records by labels in angled brackets (similar to SGML or HTML) and is easy to generate from any database software and to manipulate using simple programs or scripts. To allow coherent searching and presentation of records from both sources, fields have been mapped to MARC, and a common set of index fields is used (e.g., author/creator, title, and subject). Similar non-MARC item-level records are used for several collections in American Memory, particularly for materials that would not normally be cataloged individually in LC's catalog, such as flutes in the Dayton C. Miller collection. Several collections from awardees of the Library of Congress/Ameritech competition will be integrated into American Memory by transforming records from the awardee's database into this format, which is comparable to (and can be mapped to) the Dublin Core elements, with a small set of qualifiers. The Library of Congress

expects to adopt an XML (eXtensible Markup Language) representation for Dublin Core descriptive records in the future.

Staff from LC's Manuscript Division and Prints and Photographs Division were involved in the development of the Encoded Archival Description (EAD) standard, a document type definition for the Standard Gen-

usually added as notes rather than in subject headings or captions. Searches limited to subject fields would miss this information.

Records for some items may nevertheless contain subject terms that have not been checked against an approved thesaurus. Terms may be transcribed from earlier records or suggested by a scholar in conjunction with a special project, such as an exhibit or publication. In the Prints and Photographs Online Catalog, the uncontrolled terms are displayed separately, labeled as Topics rather than Subjects. A third category, Format, is used for terms that describe the form and genre of the original item. However, a search by subject includes all three categories. Initially, a subject search in PPOC was limited to controlled headings until reference staff confirmed that users found the distinction of little value for searching. Within American Memory, the three categories are displayed as a single group.

Users, expert and novice alike, like to look for items in archival collections by date or geographic location. For American Memory, these attributes expose many challenges since consistency across the entire heterogeneous body of materials, although clearly desirable, is hard to achieve. For published works, the date and place of publication, although often easy to ascertain, may bear little relation to the period and location described or represented, which may be of more interest to users. Fixing an unpublished pictorial work precisely in time and space is often infeasible. For most photographs, what matters is where and when the exposure was made. However, if the photographer does not record those details, precision is impossible. For unknown dates, the degree of uncertainty is conventionally represented in ways that are comprehensible to humans once a record has been retrieved—e.g., 189-?, ca. 1892, 1892 or 1893. These conventions, however, create problems for automated retrieval or sorting by date. For some American Memory collections (e.g., George Washington's correspondence), dates are recorded in a standard form so that search results can be sorted chronologically. Effective searching or sorting by date for the Prints and Photographs Division's pictorial materials proves elusive, and PPOC makes no explicit attempt to support it.

For retrieval by geographic location, a more satisfactory solution is available. The Prints and Photographs Division and the Geography and Map Division both make use of a hierarchical nation-state-county-city breakdown (e.g., United States—Illinois—Mercer County—Aledo). Whichever components are known can be recorded. This form has proved useful both for human readers and for generating lists of place-names for browsing and clickable maps that permit retrieval by state. The desirability of having certain elements within catalog records easily parsable by computer programs is reflected both in extensions to the MARC format in

recent years and in the discussions surrounding the development of the Dublin Core.

Storing and Relating Description and Content

Identifiers as Links from Description to Digital Content

The Library of Congress maintains its catalog records and finding aids independently of the digital reproductions it makes. It does not expect to embed all descriptive metadata into the content or to manage all descriptive metadata and digital content in one integrated system. This practice follows from the desire for a unified catalog that provides access to information in all forms and facilitates the sharing of cataloging effort among libraries. LC distributes copies of its catalog records to other institutions through its Catalog Distribution Service; some records include links to digital content stored at the Library of Congress. Access to overlapping content in American Memory and PPOC is supported by the same set of catalog records, but this set is separate from LC's main catalog. For the records that overlap with the main LC catalog (currently only a small proportion), copies are made and re-indexed regularly. Now LC has moved its main catalog to the new integrated library management system (ILS). The Prints and Photographs Division hopes to load the rest of its MARC records into the main cataloging system. Copies of the records will still be exported for American Memory and PPOC, for which development will continue in parallel with deployment of the ILS.

Since LC cannot predict how and where its catalog records will be used, the link between a descriptive record and a digital content "object" must be a persistent identifier in a standard format that is globally unique and, unlike Uniform Resource Locators (URLs), will not change if LC moves digital content from one computer to another. An intermediate system is needed to "resolve" the identifier to the correct physical location; when the physical location for an item changes, a record will be updated once in the resolution system rather than in the catalog record and all its copies. LC has installed and begun to use the Handle System® from the Corporation for National Research Initiatives for this purpose. Experimentation with handles as persistent identifiers in catalog records has begun for monographs and maps. The handles are based on the scheme of logical identifiers used for all materials digitized by NDLP.

Each item has a unique two-part logical identifier. As examples, dag.3g05001 is a daguerreotype portrait and musdi.139 is a reproduction of *Powell's Art of Dancing,* a dance instruction manual. Currently, the two parts of the identifier are related to names of directories and file names in the UNIX system hierarchy. In the longer term, the logical identifiers will serve more generally as unique persistent identifiers, however the content is stored. The handle for the dance manual is urn:hdl:loc.music/

musdi.139. A catalog record for this item incorporating the handle is in LC's main catalog and has been distributed to other institutions. The handle resolves to a Web-based presentation of the book, generated dynamically from its digital content, which includes page images and transcribed text. Since only Uniform Resource Locators (URLs) are usable today by most browsers and library catalog systems, the MARC record includes http://hdl.loc.gov/loc.music/musdi.139 as a URL. The use of the proxy server hdl.loc.gov provides an identifier that is usable today but is not entirely independent of physical location, since the proxy service is supplied by a particular computer. Whenever Uniform Resource Names (URNs) are deployed as a standard across the Internet, use of the proxy server can be discontinued.

Since the handles are resolved by the handle server, independently of any particular database or application, these identifiers can be used as links from any document or descriptive record. They support links from American Memory, PPOC, LC's main catalog, and from other catalogs that incorporate Library of Congress records. LC will use handles in finding aids to link to related digital content. Users, from scholars to schoolchildren, can use these handlles to turn citations in online papers into active links. The Prints and Photographs Division expects to start using handles for pictorial items now that LC has migrated its cataloging operations to a new library management system.

What Do Identifiers Identify?

An identifier, such as dag.3g05001, does not identify a single image file but a cluster of files, typically an archival master file with derivative thumbnail and service images. The number and characteristics of the images in this cluster may change over time. New thumbnail images have been generated recently for several older Prints and Photographs Division collections for better quality and more consistent sizing. In some instances, service images of a different size have been generated. Catalog records needed no changes since they contain no technical details for the image files. The metadata that expresses the relationships of the component images to the complex object (sometimes also called a meta-object), and describes the individual files, is held elsewhere. The structural information for pictorial images is currently largely implicit in file names; eventually, it will be recorded explicitly in a repository system designed to support the management of digital collections. Within American Memory or PPOC, the identifier for a picture triggers a dynamic presentation built from the associated structural metadata.

For most pictorial items, the presentation embeds a thumbnail within a bibliographic record; larger versions of the image are available by clicking on the thumbnail or a labeled link. A bibliographic record for a typical item digitized recently links to a single "picture object" with a master

image and one or two smaller service images. As mentioned earlier, the identifier in a HABS/HAER record links to a much more complex object, consisting of all image files for all the related photographs, drawings, and pages of text, with structural metadata recorded for each category of images. For the furnaces of the Sloss-Sheffield Steel and Iron Company in Birmingham, Alabama, the originals include 134 black-and-white photographs, 49 textual pages, with 20 drawings and 2 color transparencies still to be digitized. Between these extremes of complexity are the digital objects representing baseball cards; these include images of both front and back, each in several sizes.

No standard digital formats have been developed yet to represent objects as complex and varied as those created by the National Digital Library Program and other programs creating digital reproductions. Within the Library of Congress, an effort is underway to develop a common framework for the structural and administrative metadata needed for the variety of materials being converted to digital form. Early thoughts in this area contributed to the collaborative Making of America II project (http://sunsite.berkeley.edu/MOA2/), which involves five research libraries, all members of the Digital Library Federation, to which the Library of Congress also belongs. In the MOA-II project, XML is being used to represent a flexible hierarchical structure for "archival objects." Programmers from LC at the University of California, Berkeley, have developed a set of software tools to record the structural and administrative metadata, generate the XML files that represent the objects, and allow users to navigate the hierarchy and display an object's components. LC hopes to investigate whether the MOA-II archival object format and related tools might be applied effectively to its content. Meanwhile, in the commercial sector, formats for electronic books are proliferating.

When widely accepted formats exist for objects, corresponding tools or helpers for viewing, navigation, and manipulation become available. It is then possible simply to provide access to the stored objects and rely on users having the necessary tools. Individual JPEG images and PDF documents are examples of formats that can be treated this way. For more complex classes of objects, an institution that wishes to provide convenient remote access to individual works it has digitized can generate Web-accessible presentations of each object and support direct links to the presentations. This is the current approach taken by LC when it assigns handles to digitized books or maps. The handle invokes a dynamically generated presentation, in effect specifying a search for a known item within American Memory. However, the user's view of a book or a baseball card digitized at the Library of Congress may be very different from the display of similar items digitized at another institution. It remains to be seen whether users will find the variety confusing or tolerate it as a minor inconvenience given the benefits of access to a wealth of resources

distributed across the Internet. LC believes that the community will benefit from increased standardization or at least commonality of practice. Agreed formats for representing complex objects can support not only more effective interoperability but also a shared strategy for archiving and preserving digital materials for the future.

Storage Media

In 1996, the Library of Congress was faced with rapidly increasing requirements for digital storage as plans were drawn up for digitizing millions of pages of text, hundreds of thousands of pictorial items, thousands of maps and sound recordings, and hundreds of motion pictures. At the time, the cost of magnetic disk storage was $1 per megabyte. A decision was made to install a hierarchical storage management system that would automatically move less-used files to magnetic tape cartridges in a robot-controlled jukebox. Initially, a terabyte of disk storage was combined with 4 terabytes of the cheaper tape capacity (1 terabyte = 1,000,000 megabytes). All files were always accessible, but those on tape took longer to retrieve. For two years, this system proved effective but, as the number of files grew, the perceived performance dropped and backups became increasingly difficult to complete during off-peak hours. By late 1998, the cost of disk storage had dropped to between 20 and 30 percent of its 1996 cost and was expected to fall to 10 cents per megabyte by the year 2000. Use of the hierarchical storage management system has been abandoned for the time being, although LC expects to track the improvements in this class of product closely. In early 1999, LC had 19 terabytes of disk storage attached to its pool of UNIX servers and began developing an enterprise storage network that is independent of its primary data network so that backups place a minimal load on the processors and network that support users' activities. The storage network will support different media types as necessary. For example, a robotic jukebox for CD-ROMs would allow direct loading of image files delivered by scanning contractors without affecting regular network performance.

Providing Technical Support for Online Access

Indexing for Search and Retrieval

American Memory and the Prints and Photographs Online Catalog rely on InQuery (from Sovereign Hill Software, recently acquired by Dataware Technologies), a search engine, developed for indexing free text, that can recognize fields in tagged records or documents. The search system was selected for American Memory as appropriate for indexing the full text of book-length works alongside bibliographic records for multimedia materials. InQuery is not a single program but a flexible set of tools for application developers who retain complete control over visual design. It is the underlying engine for very different applications at LC,

including: THOMAS, which provides public access to current legislative information; the Handbook of Latin American Studies, a traditional bibliographic service; and LC's archive of finding aids. The ability to include heterogeneous sources of data in a single search has proven valuable for PPOC, allowing the integration of non-MARC records; integration of finding aids should be feasible in the future.

Systems for indexing free text are significantly different from those designed for indexing relational databases or traditional library catalogs. They can be configured to handle word variants automatically and to ignore punctuation. They are designed primarily to take free text as queries and return a list of documents ranked by "relevance." Each product uses its own formula for relevance, but all give higher weight to documents that contain more of the words in a query, higher still if query words are repeated in the document. Words that occur in many documents in a collection contribute less to the relevance score than words that occur infrequently. Most allow searching for phrases or for words close to each other. Strict Boolean operations, however, are seldom useful for searching free text. Initially, many LC staff familiar with traditional catalogs found the lack of Boolean capabilities in American Memory troubling and failed to notice the benefits of a system that found words anywhere, was more forgiving of inconsistency in the data, and required less exactness in query formulation by users. Lengthy undifferentiated lists of "hits" perturbed those used to systems designed when precise searching and small result sets were essential because of technological limitations. Over time, incremental changes have been made (some of which are described below) to the indexing, the search options, and the presentation of search results to address these concerns. Reference librarians in the Prints and Photographs Division involved in the design and testing of PPOC pressed for enhancements that have benefitted the users of American Memory. Most of these enhancements support more precise searching for users who wish to take advantage of the capabilities. Simultaneously, staff who had complained bitterly about shortcomings began to realize the benefits of the different indexing approach and learned how to use it to advantage. Interestingly, remote users of American Memory, whose basis for comparison is often Internet search engines, are less concerned by long lists of hits, although they do look for ways to search more precisely for what they want, particularly in the full-text materials.

When a user enters a query into the single box of a query form in PPOC or American Memory (ignoring any options), several queries are performed in the background and a combined set of results is presented. First, the query is treated as a phrase. In American Memory, the second search looks for records where all the words occur within a passage of twenty words. Since this distinction is primarily of value when searching long documents, rather than catalog records, it is omitted in PPOC. The

next search identifies records that include all the words anywhere using an implicit Boolean AND. The final search is for records that include any of the words using an implicit Boolean OR. Duplicates are removed and the three or four sets are presented as a subdivided list with entries in each grouping ranked by relevance as determined by InQuery. The explicit division of the results list was one of the enhancements introduced after complaints about long lists of hits. Users appreciate clues that help them decide how far down a hits list to bother looking. American Memory search forms now also offer the option to limit results to those that include all words entered or the exact phrase. Originally, the limit for a results set was 5,000. As the size of the resource and the volume of usage increased, the effect on performance of building unnecessarily large sets became a concern. American Memory searches now return no more than 500 records by default, although users can choose to raise the limit back to 5,000.

For the Prints and Photographs Online Catalog, the ability for users to search by creator, title, subject, and various numeric identifiers, such as call number, was considered essential. The heterogeneity of material searched within American Memory has discouraged the explicit use of field qualification, although it has been used implicitly in browsable lists of subjects (within individual collections) and in bibliographic displays where the user can click on a subject heading or name to search for other records with the same heading. Since the summer of 1998, the option to search by creator/author, title, and subject has been introduced for selected collections within American Memory. Care has been taken to include the same MARC fields in these options as are included in the corresponding options in the main LC catalog. For searching across the heterogeneous American Memory collections, the general search is still the only option provided. As the number of collections and items in American Memory has grown, the simple choice between searching all collections or within a single collection became inadequate. The ability to limit a search to collections that are primarily pictorial (or textual, cartographic, etc.) was introduced in 1997. In January 1999, American Memory introduced a new feature that permits users to pick any set of collections to search.

Under the covers, configuration options have also been changed. Full-text retrieval systems often perform automatic stemming while indexing since this results in much smaller indexes and hence faster retrieval on average. However, since a stemmed index precludes searching for exact word forms, it was determined very early during testing to be unacceptable for searching for titles in bibliographic records. By default, user queries are expanded to include word variants; users can choose to match words exactly. Another technique to reduce index size is to ignore common words, known as "stopwords." After perplexed librarians could not retrieve some titles, the initial stopword list was pruned to match the mini-

mal list used for LC's main catalog (containing only articles, conjunctions, and the most common prepositions—sixteen words in all). In making both these decisions, LC has judged precise retrieval more important than efficiency, given that the effect on performance was not obvious and the costs of computing power and disk space continue to fall.

Access Management

Some of the items digitized by the Prints and Photographs Division are subject to copyright protection or special terms of gift that prevent the Library of Congress from making them freely accessible over the Internet. As part of a prototype repository, a model has been developed for managing access for authenticated users under terms that could, when required, be specified for individual objects. However, no plans are in place for immediate implementation of such a scheme. In the meantime, access to certain collections is limited by Internet IP address to use within LC. Remote users can search and view all the records in PPOC but will not be able to retrieve some images that would be accessible if they were in the reading room.

Interacting with Users: Interface Design

Although both are accessed via Web-browsers, built with the same toolkit, and relying on much of the same code and the same indexes for overlapping content, the visual designs of PPOC and American Memory are strikingly different. Each reflects the instincts of the initial design team and their interpretations of an intended user community's expectations, balanced by technical considerations and modified by incremental changes made in response to feedback. PPOC is plain, with extensive textual explanations and a look reminiscent of the popular text-only catalog displays used previously in the reading room. Initially, reference staff had pushed for a page-oriented display with Next Page buttons, concerned that a scroll bar would be difficult for users to manipulate; this had to be abandoned after it proved too difficult to calculate where page-breaks would come when images were of different sizes. Some of the professional picture researchers who have used the reading room for many years have been slow to adapt to new technology; for most users, however, the mouse is either familiar or soon mastered. American Memory uses color, space, and icons more freely, expecting its primary users to know how to scroll and click through Web pages. PPOC gives priority on displays to information about accessing originals and ordering reproductions, whereas the focus for American Memory is on the online content. However, the underlying functionality of the two search systems is the same since the organization of the descriptive records and digital content and the search capabilities are identical. Each system is constructed of four layered components as shown in Figure 2,

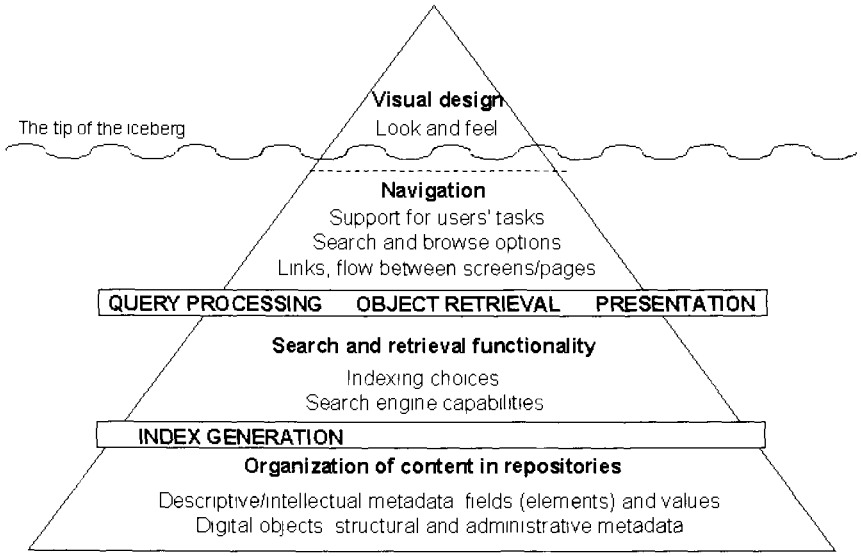


Figure 2. Search Interfaces Rely on Underlying Content and Indexing Capabilities.

a model that describes how many of LC's search and retrieval applications are constructed. They share the three lower layers, differing only at the visual design level.

As development of the two interfaces has continued, staff in the Prints and Photographs Division and the National Digital Library Programs have learned that there are advantages to increased commonality where feasible, if only because the single programming team is able to make enhancements more quickly. Although the bibliographic displays are different, the thumbnail grids and page-turning displays are generated by the same code. Ideas from either community feed the iterative development of both interfaces.

The public mission of the Library of Congress and the requirements of the Americans with Disabilities Act mean that both PPOC and American Memory make less use of icons and graphics than many of today's Web pages. Fast loading of pages is essential to reach the widest possible audience including schools and homes with slow network connections and older computers and software. Pages must be constructed in a way that is usable by assistive software for the handicapped. Rather than create text-only pages as alternatives, LC has chosen to be conservative about adopting technology that may make its collections less accessible. As an example, LC does not use frames in its page-turning presentation.

Facilitating Use of the Images

Two factors control re-use of images: whether a copy is available in a form suitable for the intended use, and whether re-use is permitted by copyright law. Roughly one-third of the electronic mail that comes into the reference desk of the Prints and Photographs Division relates to rights and reproductions. For the American Memory Help Desk, the proportion is much less—around 5 percent. As an agency of the U.S. Government, the Library of Congress claims no rights in the digital versions of the materials it converts. However, rights relating to the original materials pertain also to derivative works, including digital reproductions. Since many of the photographs in the P & P collections were not published and may have been created for hire, ascertaining the copyright status and the identity of possible rights holders for individual items is usually infeasible. For public access, LC has focused on converting materials produced by the U.S. Government, those likely to be out of copyright by virtue of their date of creation, or collections where a single organization or individual appears to hold copyright and commercial interest is unlikely. However, by making items accessible through American Memory, the Library of Congress does not (and is not legally empowered to) warrant them to be in the public domain. When contact has been made with copyright holders, they have usually been happy to allow educational and research use, and LC records these permissions on catalog records and on the Web pages introducing collections. Other forms of right may also apply, including privacy for the individuals represented in photographs. Whether a particular re-use is permissible under the "fair use" doctrine or violates rights is not clearly defined by law or regulation. U.S. law only specifies factors to be considered by a court adjudicating a case brought by a rightsholder. Since LC claims no rights itself and is prohibited from offering legal advice, requests for permission to re-use cannot simply be answered "Yes"—to the frustration of librarians and users alike. However, LC makes every effort to describe its understanding of the factual background and known rights for each collection in order to assist users.

Users who visit the Prints and Photographs reading room are often seeking pictures to use in publications. Until recently, only photographic reproductions could provide publication quality copies. LC's Photoduplication Service offers photographic reproductions for a fee set to cover costs. Prints are made from copy negatives or interpositives; when a reproduction request is received for an item for which no copy exists, one is made. Remote access has generated a growing volume of interest in reproductions for personal use—e.g., when family members are recognized. The archivist of the Institute of Regional Studies at North Dakota State University reports a steady stream of requests for reproductions of photographs from the institute's collections in American Memory (The Northern Great Plains:1880-1920). Requests come not only from those with ties

to the area but from people who simply like the images of pioneer life. Users with appropriate equipment and fast enough network connections can download the images directly. However, the Prints and Photographs Division reference librarians find that few users are yet comfortable downloading large files; most still prefer to wait for a traditional photographic reproduction. In the future, the Photoduplication Service may use high-quality scans from original negatives (as made for the HABS/HAER collections) as the source for prints of photographic quality.

Reference Services

Reference staff in the Prints and Photographs reading room have seen online access to images as reference surrogates as a goal for many years. They have pressed for convenient one-stop access for both staff and patrons to the division's holdings. Compared with earlier catalog systems, the current Prints and Photographs Online Catalog has advantages beyond a growing volume of content and the ability to serve remote users. As mentioned earlier, free text search capabilities compensate for less than perfect descriptive data. Simultaneous access to images from any workstation with a Web browser creates new possibilities for productivity by both staff and patrons. Recently, many thousands of uncataloged copy negatives were added to PPOC; skeletal records have a reproduction number and a link to corresponding digital image files but no title, creator, or subject headings. Since, following LC's recommendation, reproduction numbers are often cited in publications, staff and patrons can benefit from the ability to retrieve and view images even if they have not been cataloged.

Since PPOC has been accessible over the Internet, the volume of electronic mail reference questions to P & P has doubled from thirty-five messages per month to seventy. Many are basic questions to which answers are already available online, including those seeking permission to use the images or to obtain reproductions. A direct link has been added from each bibliographic display in PPOC to a page that cites the reproduction number and provides detailed information on the services provided by the Photoduplication Service and general information regarding copyright and other potential restrictions on use. The American Memory Help Desk provides standard responses to common questions in a list of Frequently Asked Questions.

Demand for reference service in the P & P reading room has always been high. Remote online access to images, through American Memory and PPOC, seems neither to have stimulated use of the physical collection nor to have reduced the number of users visiting the reading room. Reference staff, however, have noticed that the number of free-lance professional picture researchers using the collection has increased, suggesting that remote access to images over the Internet has either stimulated growth

in this niche service industry or allowed individual researchers to discover and explore more sources for pictorial material for their customers. Another effect of online access to images is the need for increased communication among divisions within LC and with other institutions about the availability of digital reproductions for online viewing or use. Effort has been needed to ensure that electronic mail questions on pictorial materials (or any other specialized topic) are dealt with consistently, whether sent to the American Memory Help Desk, LC's main Web service, or directly to a division.

Conclusion

The challenge of providing effective online access to visual materials goes far beyond the process of digitization. Indeed, the Library of Congress' experience suggests that by using consultants and contractors expert in the physics inherent in scanners and output devices, the mathematics of image transformations, and schemes for color management, and with the engineering skill to build systems that control them, it is possible to develop procedures that provide high quality images in large volumes. Harder to resolve are the underlying problems of organizing and describing visual materials cost-effectively (whether digitized or not) in ways that help users find the information, illustration, or evidence that they need. Practices developed for the published printed literature need considerable adaptation. Experience with the large FSA-OWI and HABS/HAER collections suggests directions that online browsing, treatment of pictures in groups, and free text retrieval can reduce the need for full cataloging of individual items.

Distinctions between the human interface presented by digital images on a computer monitor and by a folder of prints on a large table provide another aspect of the challenge. Enhancements in interfaces will be inevitable but gradual. Grids of thumbnails provide one approach for side-by-side comparison and rapid browsing through groups of images. Technological advances in processors, networks, and monitors will handle images faster and better and provide more options. More generally, approaches to interface design will improve through better understanding of how people interact with visual materials. At the same time, human ingenuity will suggest new ways to cope with or take advantage of characteristics of the online environment. Consider the use of microfilm: few users are enthusiastic about the medium, but experienced researchers become skilled at scrolling rapidly through a reel focusing on distinctive patterns (such as "target" pages) to indicate when to stop. The modular design of the architecture that supports American Memory and the Prints and Photographs Online Catalog will permit incremental enhancements in response to changes in the technical environment and the expectations of users.

An important objective for the Library of Congress is enhanced access to its comprehensive collections. LC has already made digital reproductions of hundreds of thousands of photographs. However, these constitute only a tiny fraction of its pictorial resources. LC seeks cost-effective solutions to provide integrated access to resources in many forms of expression, whether digitized or not. LC staff most closely involved with developing the current practices for providing online access to pictorial material are hesitant to call them best practices, preferring to consider them as appropriate practices given the state of technology and the Library of Congress' institutional objectives and constraints.

REFERENCES

Betz, F. W. (1982). *Graphic materials: Rules for describing original items and historical collections.* Washington, DC: Library of Congress. Retrieved September 3, 1999 from the World Wide Web: http://www.TLCdelivers.com/tlc/crs/grph0199.htm.

Betz Parker, E. W. (1985). The Library of Congress non-print optical disk pilot program. *Information Technology and Libraries, 4*(4), 289-299.

Flynn, M., & Zinkham, H. (1995). The MARC format and electronic reference image: Experience from the Library of Congress Prints and Photographs Division. *Visual Resources, 11*(1), 47-70.

Frey, F. S., & Reilly, J. M. (1998). *Digital imaging for photographic collections: Foundations for technical standards.* Unpublished report by the Image Permanence Institute, Rochester Institute of Technology. Final Report to the Office of Preservation, National Endowment for the Humanities (NEH GRANT PS-21084-95).

Kenney, A. R.; Shapiro, L. H.; with Berger, B.; Crowhurst, R.; Ott, D. M.; & Quirk, A. (1999). *Illustrated book study: Digital conversion requirements of printed illustrations.* Retrieved October 27, 1999 from the World Wide Web: http://www.library.cornell.edu/preservation/ill_bk_cover.htm.

Library of Congress. (1995a). *Thesaurus for graphic materials: I. Subject terms.* Retrieved September 3, 1999 from the World Wide Web: http://lcweb.loc.gov/rr/print/tgm1.

Library of Congress. (1995b). *Thesaurus for graphic materials: II. Genre and physical characteristic terms.* Retrieved September 3, 1999 from the World Wide Web: http://lcweb.loc.gov/rr/print/tgm2.

Library of Congress. (1997). *Request for proposals for digital images of pictorial materials.* Retrieved September 3, 1999 from the World Wide Web: http://memory.loc.gov/ammem/prpsal9/coverpag.html.

Library of Congress. (1996). *Cataloger's desktop* [CD-ROM, updated annually]. Washington, DC: Library of Congress Cataloging Distribution Service.

Ostrow, S. E. (1998). *Digitizing historical pictorial collections for the Internet.* Council on Library and Information Resources. Retrieved September 3, 1999 from the World Wide Web: http://www.clir.org/pubs/reports/ostrow/pub71.html.

Reilly, J. M. (1995). Technical choices in digital imaging: The technical images test project in review. In P. A. McClung (Ed.), *RLG Digital Image Access Project* (Proceedings from an RLG symposium held March 31-April 1, 1995, Palo Alto, CA) (pp. 85-93). Mountain View, CA: Research Libraries Group.

Acronyms

DLF	Digital Library Federation
FSA	Farm Security Administration
G & M	Geography and Maps Division, Library of Congress
HABS	Historic American Buildings Survey
HAER	Historic American Engineering Record
LC	Library of Congress
NDLP	National Digital Library Program, Library of Congress
OWI	Office of War Information
P & P	Prints and Photographs Division, Library of Congress
PPOC	Prints and Photographs Online Catalog (pronounced pea-pock)
URL	Uniform Resource Locator
URN	Uniform Resource Name
XML	eXtensible Markup Language

Recent Developments in Cultural Heritage Image Databases: Directions for User-Centered Design

CHRISTIE STEPHENSON

ABSTRACT

FROM 1995 THROUGH 1997, SEVEN CULTURAL HERITAGE repositories and seven universities collaborated on an extensive demonstration project called the Museum Educational Site Licensing Project (MESL) to explore the administrative, technical, and pedagogical issues involved in making digital museum images and information available to educational audiences. This article reviews the MESL project's methods and findings in a number of areas—descriptive metadata, database design, interface design, and tools for use. It discusses more recent development efforts in extending the model for digital image delivery of visual resources to higher education audiences. Finally, it suggests how to proceed by posing a number of user-centered questions about the design goals for networked access to the vast visual resources of the cultural heritage community. Selected projects from the literature of computer and information science are discussed to stimulate thinking about avenues for research and to focus project design goals.

INTRODUCTION

In his 1996 review article, "Image Databases: The First Decade, the Present, and the Future," Howard Besser (1997b) presented an overview of ten years in the development of image databases designed to provide access to cultural heritage information. How much further have image delivery systems progressed in the past several years of rapid technological change? This article examines the Museum Educational Site Licensing

Christie Stephenson, University Library, Digital Library Inst., 300-B Hatcher North, 1205, University of Michigan, Ann Arbor, MI 48109
LIBRARY TRENDS, Vol. 48, No. 2, Fall 1999, pp. 410-437

Project as well as several more recent projects representing the current state of development of cultural heritage image databases for academic use. It reviews recent literature in areas such as image indexing and retrieval, interface design, and tool development, and urges a reexamination of our efforts in those areas based on a more rigorous analysis of user needs.

The Museum Educational Site Licensing Project (MESL)

In 1995, a boldly envisioned demonstration project was launched which provided new insights into the issues of building large-scale image databases for the delivery of cultural heritage information to higher education audiences. The Museum Educational Site Licensing Project was a collaborative project, involving seven universities and seven cultural heritage repositories, to investigate the administrative, legal, economic, technical, and educational issues involved in providing networked distribution of museum content for educational use. During the two years of the project, the seven museums provided nearly 10,000 digital images and accompanying descriptive metadata records. These were distributed to the seven universities, each of which developed their own local delivery system. Although much of the project focused on the legal and administrative issues of licensing, it also provided a valuable testbed for exploring a range of issues related to the building and subsequent use of large image databases from disparate collections of cultural heritage images and data. Although only limited rigorous research was undertaken within the brief duration of the project (1995-1997), the project participants were able to report a number of useful observations about descriptive metadata, database and system design, interface design, and tools for use of the images and information (Stephenson & McClung, 1998).

Descriptive Metadata

While traditional analog visual resource collections in educational institutions have depended on physical arrangement and local cataloging to provide access, the descriptive metadata provided with the MESL images was extracted from data that already existed in the museum collection management systems–legacy data from systems built primarily to handle the internal informational requirements of the repositories. The first step in the process of providing useful descriptive metadata to the universities was to agree on a common metadata structure to which the individual institutions could map their own data. The MESL Data Dictionary, defining records composed of thirty-two data fields, was developed to serve this purpose. The museums mapped their data to this structure and developed export routines to extract data from their collection management systems and populate the MESL data records. If they did not

have data for a specific field, it was left blank. The completeness of the records varied both within a single institution as well as from institution to institution, depending on the level of documentation any object might have received.

The data populating the various fields in the records were not standardized in any way. Because museums have only recently begun to adopt principles such as authority control, there were many variations in data values for standard entries such as artist names. And very few of the museums had supplied any subject access beyond the most general terms; when present, they were inconsistently applied. The museum data had been created for collection management, not public access; making it available to a new user group, beyond the museum staff for which it was created, revealed its inconsistency and limited usefulness for open ended searching. One can postulate that what was true for the MESL museum data is generalizable to most museum collections' management information (Dowden, 1998).

Database Design

Each of the universities participating in the MESL project designed its own delivery system using local resources and frequently building on existing information systems and infrastructure. A variety of backend databases were used, ranging from Filemaker Pro and Microsoft Access to OpenText (Besser, 1998). This heterogeneity of the system design effort at the universities not only led to differences in the look and feel of the interfaces but also to somewhat surprising differences in the search results when the same queries were posed to each system (e.g., see Figure 1). Besser (1997a) reports on these sometimes dramatic anomalies. They resulted from local decisions about what to index as well as characteristics of the local search engines at each of the sites. Some institutions decided to index only selected data fields while others provided full-text searching across all data (including unstructured full text) as well as field-specific searching. In addition, a number of the search engines handled functions such as phrase searching, truncation, and stemming differently and in ways that were frequently not apparent to users.

The entire MESL data set was mounted locally by each of the participating universities rather than served from a single central distribution point. Each university mounted the MESL data set as a separate database, even when they had other image databases available to their users. Though a number of them expressed the desire to integrate MESL data with image resources from other disparate sources, the limited life span of the MESL project made this infeasible. The experience of mounting the MESL data gave the universities insights into the challenges they would face in such an undertaking. Since the conclusion of the project, several of the participating universities have made strides in this area.

Figure 1. Searches on **card table** in the MESL databases of the University of Virginia (top) and the University of Illinois (bottom). The search at Virginia retrieved 4 records; at Illinois, only 3 records were retrieved. Note that the Virginia system displays multiple views of the same object.

Interface Design

In addition to creating search and retrieval systems for the MESL images and data, each of the participating universities designed their own search and browsing interfaces. Despite independent development, most of the interfaces were very similar in look and feel. The Web browser window provided the basic interface. Standard Web forms were used to present search options. Users could choose between a simple search and complex or Boolean searches and could search one museum collection or across all of the collections. Pull down menus allowed users to specify a particular field to search or they could search all indexes. Each of the universities implemented more or less standard ways of displaying search results within the browser window. For browsing, a grid of thumbnails with brief identifying captions was the most typical presentation; if many thumbnails were returned in response to a query, users had to page through multiple screens. Users could also select a brief record display where screen elements included fielded textual data presented in something like standard bibliographic format with the associated image or images next to it. They also were given the option to view larger versions of the images as well as full textual records showing all data supplied for an object. The University of Virginia implemented a search results display option that returned unlabeled thumbnails, giving the user the ability to scan and mark thumbnails as an initial visual interface for making relevancy judgments (Besser, 1998) (see Figure 2).

Tools for Use

Most of the functionality provided by the MESL university participants was constrained by their choice of the Web as a delivery mechanism. The state of Web development at the time as well as local limits on available technology support precluded the development of additional functionality such as Java-based tools. Faculty and student users searched the database and used cut-and-paste methods to create class Web pages or include images and descriptive text in papers or presentations. At the University of Virginia, very simple templates were developed to facilitate the creation of side-by-side image comparisons and online Web exhibitions (http://jefferson.village.virginia.edu/inote/index.html), but they, too, depended entirely on the use of cut-and-paste methods (see Figure 3).

At the University of Maryland, however, a variety of factors contributed to the development of a sophisticated software product which simulated the function of a slide library's light table. It supported faculty members in the process of selecting, organizing, and arranging material for delivery in the classroom, mimicking the side-by-side projector environment typically used in teaching art history. Maryland's delivery system, now known as ISIS (Interactive System for Image Searching), was developed by a team of programmers, instructional designers, and the faculty

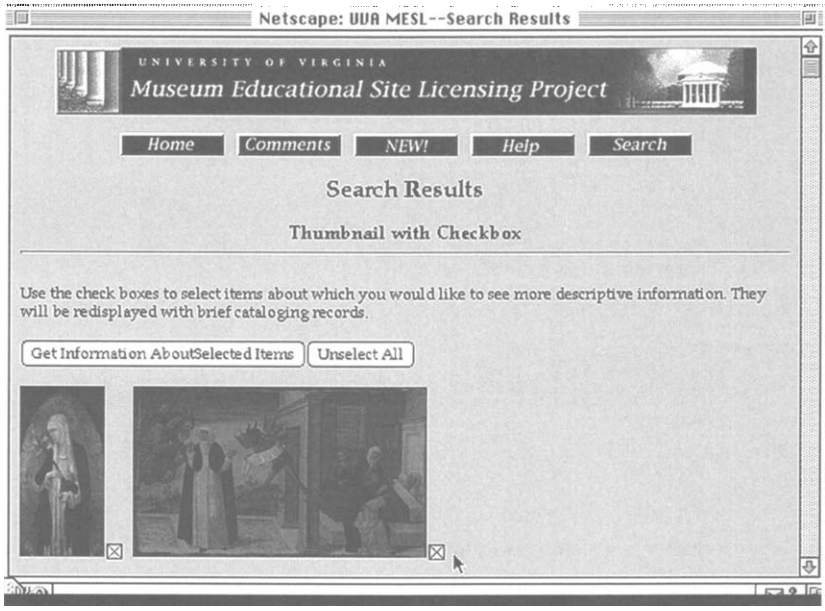


Figure 2. University of Virginia MESL displays—Thumbnail with Checkbox (no captions) and Thumbnail with Brief Record

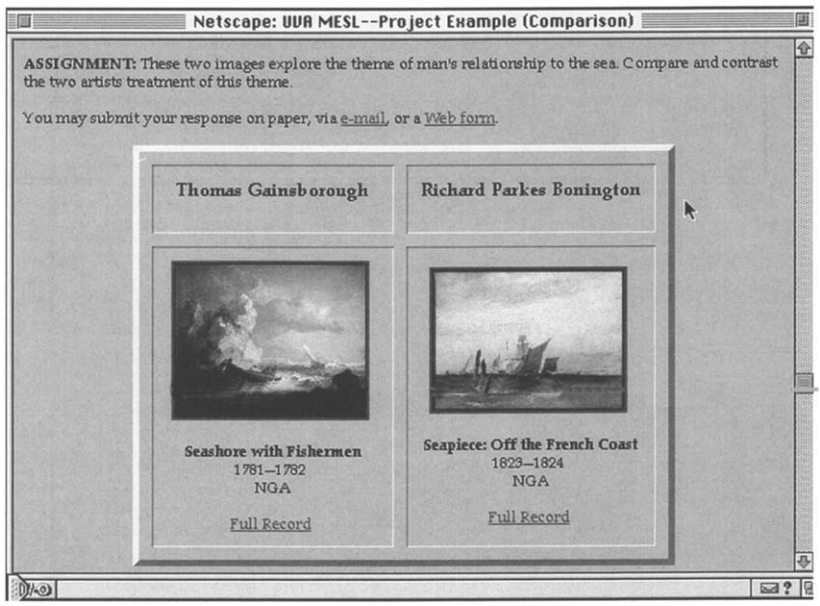


Figure 3. University of Virginia, MESL Comparison Template.

members themselves through an iterative process that continued throughout the duration of the MESL project (Borkowski & Hays, 1998) (see Figure 4). The commitment to this product resulted directly from the early organizational decision to base MESL development in the art department, thereby involving end users in design decisions from the outset. The impact it had on the success of the MESL project at Maryland was remarkable (Promey, 1998).

Recent Federation and Expansion Efforts

Since the end of the MESL project in July 1997, efforts have been underway on a number of campuses to provide federated access to diverse image collections, allowing users to search individual or multiple repositories from a single search interface. This development is the next step in the effort to provide users with broad access to information about cultural heritage objects held locally as well as those licensed or otherwise made available from other sources. Projects at the University of Michigan Library and Harvard University Museums and Libraries serve as representative examples of these undertakings. In addition, the work begun in the MESL project is being continued and expanded by AMICO, the Art Museum Image Consortium, working in cooperation with the Research Libraries Group (RLG).

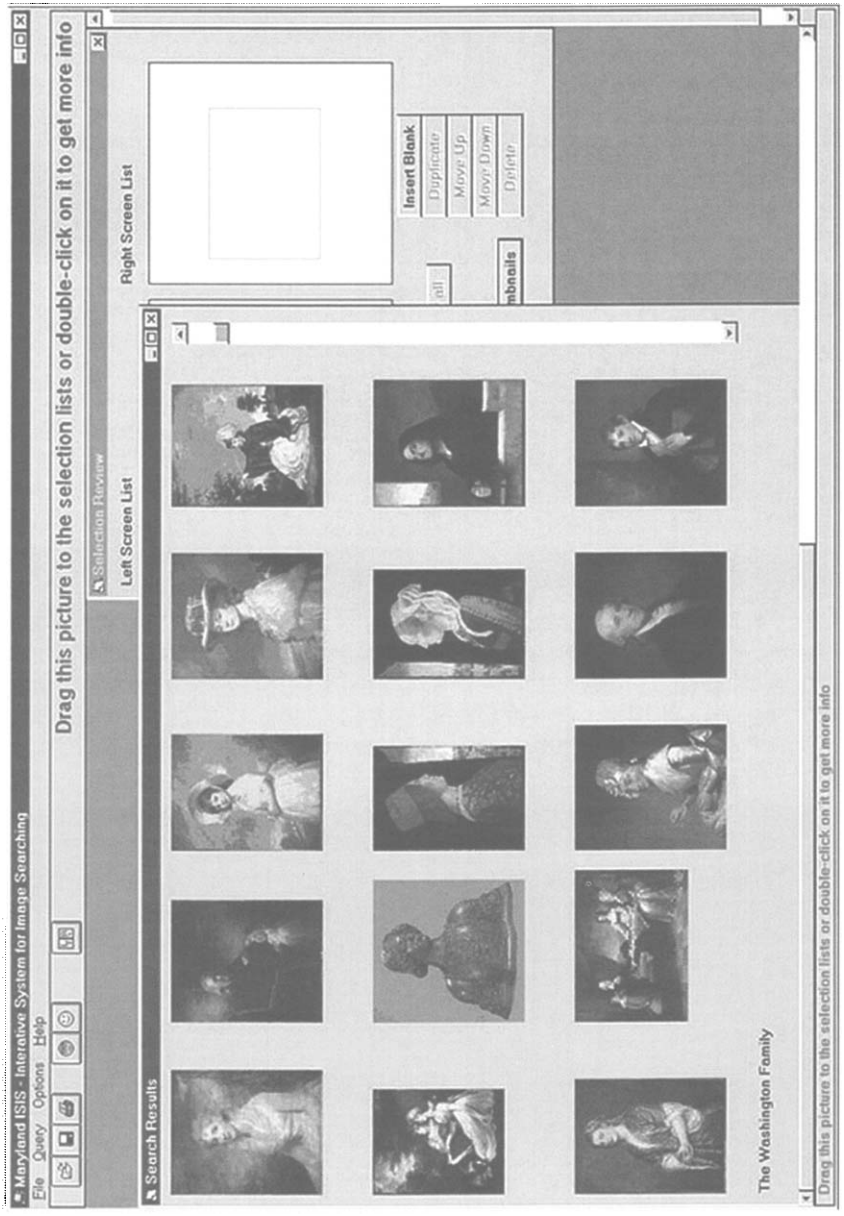


Figure 4. Maryland ISIS: Image Selection Screen.

University of Michigan

In 1997, the University of Michigan Library began to create an architecture for federating access to image databases through the Image Services component of its Digital Library Production Service (DLPS) unit. Among their stated goals, they seek to "provide to faculty, staff, and departments a standardized, base level, extensible architecture for putting images online" (http://images.umdl.umich.edu/dlps/is.html). The staff of the Image Services group established a core metadata set for visual images by analyzing existing metadata schemes. The Michigan Image Access System merges data drawn from the collection management databases of a number of campus visual resource and museum collections, as well as that provided with licensed image collections. Data are extracted from each of the separate management databases, mapped to the shared metadata scheme, marked up in SGML, and indexed using OpenText. Users are provided with the option of searching any individual collection by the metadata elements in its own data or multiple databases by a more limited number of common metadata elements.

The system design is based on the fundamental assumption that image or object databases are created to meet the management needs of the particular repositories and should remain independent from the provision of public access. Standardization of public access through the DLPS provides consistent service to meet instructional and research needs. Image Services provides a standard interface for searching all collections as well as a search form customized to each collection (see Figure 5). They have articulated a set of incremental improvements to the system, including the ability to create "personal collections" (http://images.umdl.umich.edu/info/arch/arch_summ.html).

Harvard University

Harvard University has over 8 million objects and images in its libraries, archives, and museums. Its diverse institutional environment is a challenging testbed for providing integrated access to cultural resources collections. Over the past two years, representatives from museums, libraries, and archives at Harvard and Radcliffe, working together with the Library Office of Information Systems, have been engaged in a process to build a shared union catalog of visual resources. The goal of the union catalog project is to create a common database where users can discover Harvard's wealth of visual resources and be directed to the holding repository for more detailed information or access to materials.

To date, this project, known as Visual Image Access (VIA), has devoted much of its effort to the process of agreeing on a common data structure for its union catalog as well as reaching consensus on the scope and functionality of the catalog. In the first phase of the project, currently underway, object and collection records from six diverse collections

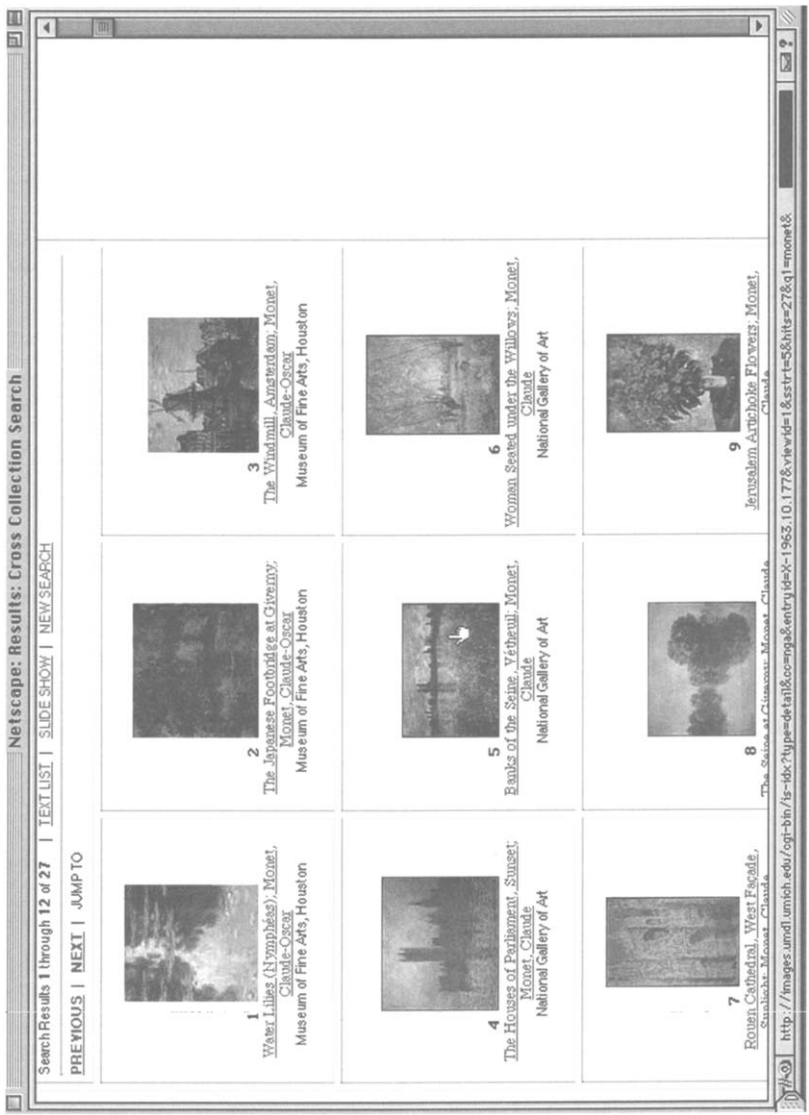


Figure 5. University of Michigan Library Image Services; cross collection multiple image retrieval display.

will be merged into a single database; digital images may be associated with the records but are not required. At the outset, VIA will include only cultural heritage materials based on the existence of similar metadata structures and distinct functionality required for use. Like the Michigan Image Server Program, VIA acknowledges the separate and primary functions

of each repository's collection management or access system; it does not intend to dictate local practice but to federate local records. And as at Michigan, the intent is to provide access not only to locally held images or objects but to licensed image content as well. Implementation of the VIA system was scheduled to begin in early 1999 (http://sylvia.harvard.edu/~robin/viascope.htm).

Art Museum Image Consortium (AMICO)

The Art Museum Image Consortium, founded in 1997, is in many respects the successor to the MESL Project. In its project description, AMICO is characterized as "a not-for-profit consortium dedicated to creating a digital library" documenting the collections of its members and making that library available for educational use. It currently includes over 20,000 images and object records from over twenty contributing institutions and anticipates a growth rate of about 50,000 objects a year (http://www.amn.org/AMICO/). At present, those resources are being distributed to twenty universities participating in a year-long testbed project. Unlike MESL, where all the data were distributed to each of the participating universities, AMICO is currently providing centralized distribution to the testbed participants through the Research Libraries Group. As in MESL, the AMICO data set uses existing collections management information extracted from the contributors' systems and mapped to a common data dictionary. The data set and images are made available to AMICO testbed users through a modified version of RLG's Eureka interface (see Figure 6) (http://www.rlg.org/amicolib.html).

Although each of these projects has slightly differing goals, all of them attempt to provide unified access to images and information from diverse collections. In time, each may develop or adopt a system architecture that facilitates network-distributed discovery. However, at present, all are still merging data locally. Though they incorporate more sophisticated design elements than the MESL delivery systems, the underlying data are similar in their lack of consistency, and the interfaces are quite similar to those developed by the MESL participants. Each of these systems will likely challenge and frustrate users in many of the same ways that the MESL implementations did.

Understanding User Needs and Expectations: Next Steps

In looking ahead, it seems clear that there are numerous obstacles to overcome in order to realize our ambitious goals for digital image delivery systems. While some of these lie clearly in the realm of technology, many depend on collaboration between human-computer interaction specialists, librarians and collection managers, evaluation specialists, and end users, both sophisticated and naïve. In his 1996 article, Howard Besser

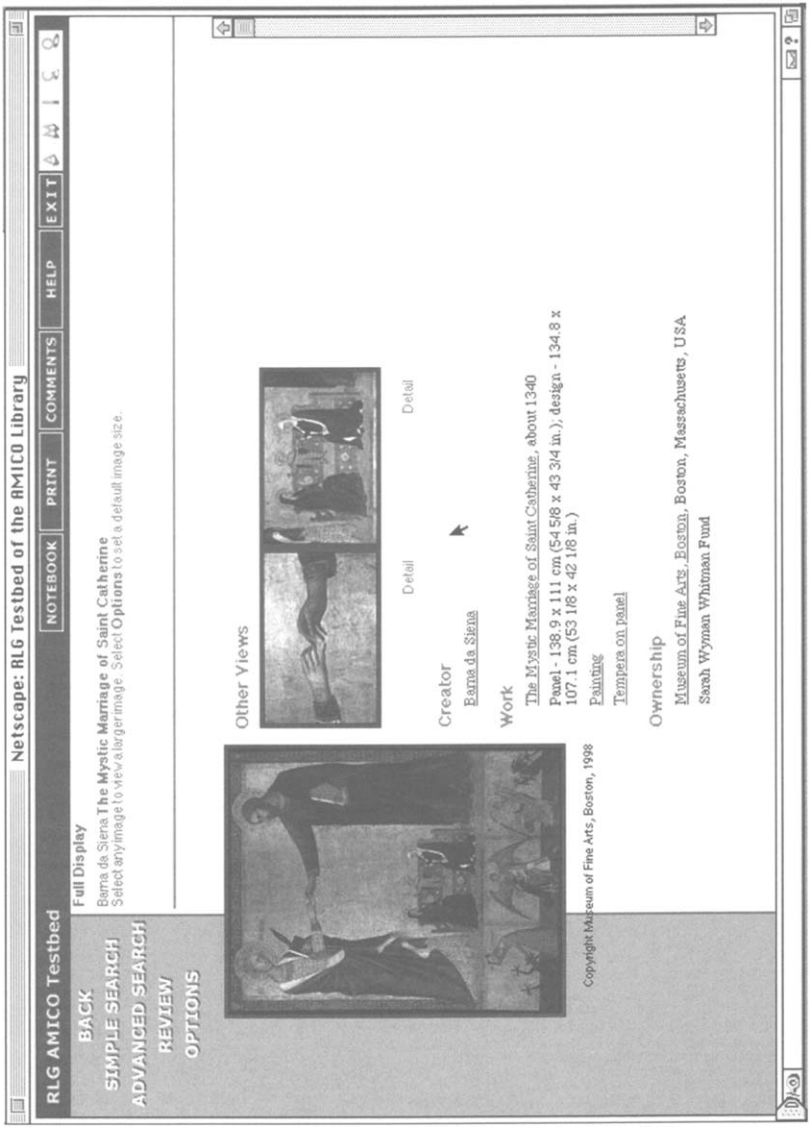


Figure 6. Sample Screen from the RLG AMICO/Eureka Interface.

(1997b) focused his articulation of next steps in a number of technical areas: preservation, authenticity, and integrity of information; image standards; image quality issues; and retrieval. Rather than revisit and reassess our progress on each of those issues during the intervening years, it may be useful to look ahead through a different lens—one that puts user needs

and expectations, rather than technology, at the fore. Although the creation of image rich digital resources certainly represents a series of technological challenges, it is critical to give adequate attention as well to the fundamental questions of audience, user behavior, and use.

The examination of several broad questions can assist in developing a user-based model for directing development of image delivery systems and guiding future research:

- For whom are we building our image delivery system?
- What is it that we are building and for what purposes do those users want to use it?
- What functionality do our users need to use what we build?

This kind of design model, called User-Task-System or U-T-S (Lindermeier & Stein, 1991) helps ensure that system design does not limit users and uses. Instead, a thorough analysis of user characteristics and requirements will drive sound system design.

Defining the Primary User Group

Moving from the analog to the digital world, it becomes increasingly difficult to characterize the users of our image delivery systems. In the past, it was possible to know much about our users by restricting physical access to collections to a specific group or requiring registration prior to use. The closest parallel in the digital world is to allow access only from a specific set of workstations. But as a general rule, one of our overriding principles in the digital realm is to *extend* access, not restrict it. This means our systems are likely to serve both traditional and new users of image resources; local and remote users; sophisticated and naïve computer users; users supported by on-site assistance and who will interact unmediated with our systems; children and adult learners; and so on.

In the MESL project, the difficulty in serving these diverse user groups effectively was demonstrated even when access was limited to a specific university community. Traditional users of images, primarily art history students and faculty members, were frustrated by the absence of particular works of art. Nontraditional image users were often stymied by the lack of subject access to the works in the database. Both groups were sometimes frustrated by the limited functionality of the delivery systems that relied on relatively simple Web design.

At least in the short term, it is unlikely that systems can be built that serve all user groups equally effectively. Ideally, good system design would isolate digital objects in a repository, and any number of front-ends could be customized for specialized user groups. However, the initial design effort is likely to focus on a specific set of users and uses. Even if the resource is aimed at a broadly defined generic user group and a relatively

use-neutral presentation such as search and retrieval, it is possible to define some basic characteristics of the user population. As a part of a rigorous planning process, the characteristics of the intended primary user group can and should be articulated and assumptions about them enumerated. Mechanisms to test those assumptions can be built into the iterative system design process. Those mechanisms might include a variety of quantitative and qualitative techniques including log analysis, online user surveys, usability studies, interviews, and focus groups. By focusing on the needs of clearly defined user groups, it will be possible to better understand system requirements, better target development efforts, and more reliably test design decisions.

One outstanding example of such a model is embodied in the work undertaken by a research team from the University of Maryland, working with the Library of Congress National Digital Library Program (Marchionini et al., 1998). The goal of the research was to develop interfaces for NDL content (much of it consisting of visual images) "guided by an assessment of user needs and aimed to maximize interaction with primary resources and support both browsing and analytical search strategies" (p. 535). The project consisted of several phases: problem identification and team development, interface design and prototyping, and tool development. The Maryland team emphasizes the importance of developing and testing principles and guidelines for user-centered iterative design for delivering digital library content to a variety of end-user communities (p. 553).

Understanding Anticipated Uses

Once the primary user group or groups are explicitly described, the next challenge is to articulate the range of uses that need to be supported and to understand the implications for system and interface design. In the higher education setting, local image collections have primarily consisted of collections of surrogates, usually slides, built to serve a curriculum support function. In addition, library special collections departments and museums have built collections of images or objects. The functional roles of these collections may be less clearly articulated than those of the visual resources collection, making it more challenging to develop appropriate design criteria for delivery and use.

As we begin to digitize these collections, a number of questions must be confronted. What are our goals as we build digital image delivery systems? Are we replacing local slide libraries with digital image collections with curriculum support as their primary goal? Perhaps we are creating collections of document surrogates, with item level description, to obviate the need for handling precious or fragile originals. Or are we digitizing quantities of images for which we will never be able to provide item level description to facilitate access to underused collections? Are we building union catalogs of records about objects and image collections, primarily

as location devices, leading users to repositories and originals? Are we licensing image databases and making them available as we would our many bibliographic databases as a use-neutral electronic resource? Or are we building hybrid systems, merging collections created for some or all of these purposes, into large supersets, where audience and aims become increasingly indeterminate?

Depending on the answers to these questions, it is possible to begin to articulate the range of functionality needed to discover, retrieve, and use images in a collection and to focus development on supporting the stated system goals. In the case of the MESL project, the product was arguably a hybrid collection built to explore a range of issues rather than to serve a specific articulated goal. The project sought to document and assess the ways in which images and their associated information were used to further our understanding of searching strategies, image quality needs, user tolerance levels, and adequacy of access vocabularies. The participants were also committed to understanding the system requirements necessary to facilitate pedagogical uses of the images and information. Although unable to conduct detailed log analysis and other focused research, the participants did engage these questions and report rich anecdotal evidence about a number of them. In order to build on the MESL experience, it is useful to examine our findings in relation to selected current research in computer and information science. This review suggests additional avenues of exploration which may help future system developers to more successfully deliver the necessary functionality to end users.

Supporting Discovery and Retrieval

In hybrid systems such as MESL, the Michigan federated image server, or a distributed networked delivery system, what are the requirements that must be met to support discovery and retrieval? Writing in 1995, Hinda Sklar (1995) articulated three of the elements of basic functionality for image databases. These are: (1) to perform a range of searches, formulating both simple and complex queries, and searching using both controlled vocabularies and keywords; (2) to search many collections in a single search from one location; and (3) to discover unknown resources. Achieving this functionality depends on descriptive metadata structure and data interchange architecture, metadata values, the search and retrieval system itself, and the interaction of the user and that system.

Metadata Structure and Data Interchange

A thoroughgoing discussion of the current state of emerging descriptive metadata and data interchange standards for images is beyond the scope of this article. There are numerous testbed projects underway focusing on metadata and interoperability requirements, particularly the Dublin

Core and Z39.50 development work undertaken by the Consortium for the Interchange of Museum Information (CIMI).

Metadata Values

Much anecdotal evidence was gathered in the course of the MESL project to indicate that both the lack of descriptive metadata for subject access and the lack of standardization across existing metadata values will continue to frustrate users trying to locate images. Although museums are beginning to recognize the need for standards to facilitate information interchange, the paucity of controlled access points will continue to adversely affect certain uses of our growing image databases. And large numbers of images will never be described at the item level, further frustrating users.

Rasmussen, writing in 1997, reviews the growing body of research addressing users of images and image databases, query analysis, and user needs and behaviors. A few of those projects are highlighted here. Since strategies must be developed for dealing with the lack of descriptive metadata for image access, even more effort could be made to incorporate and extend research on the behaviors of image seekers and image indexing practice. By better understanding how images are sought, it should be possible to prioritize efforts to augment subject access in the ways that will have the greatest impact on user satisfaction.

Query Analysis

There have been surprisingly few studies on user queries. Rasmussen (1997) reviews several query analysis projects from the early 1990s. Several more recent projects include that of Collins (1998), who studied user queries in two historical photographic collections. She found that generic subject terms appeared most often in those queries followed by terms referring to time and place. Armitage and Enser (1997) collected queries from seven picture collections and sought to develop a general purpose schema for categorizing user requests for images. Janney and Sledge (http://www.cimi.org/documents/z3950_app_profile_0995.html) analyzed 1,500 queries made in museums (not necessarily image queries) in order to develop an attribute set for information retrieval for museum information. Hastings (1995) studied and categorized queries of art historians working with a database of digitized images and associated text. Additional query analysis is fundamental to understanding and improving access to images.

Evaluating Indexing Methods

Having acquired a better understanding of query structure, the next step is to evaluate the effectiveness of image indexing in answering those queries. Again, there seems to be little empirical research in this area. A number of authors make a strong case for additional quantitative research

on the effectiveness of various indexing methods (e.g., Layne, 1994; Tibbo, 1994). For instance, Layne (1994) makes a case for identifying and indexing attributes which provide groupings of images rather than access to individual images, allowing the user to visually scan results to make comparisons or relevance judgments. Jorgensen (1998) recommends that "assumptions underlying controlled vocabularies and newer descriptive tools...should be tested [and] that new ways of indexing images would perhaps improve the process of image retrieval" (p. 171). In her research, she found a disjunction between user image-seeking behavior and current image indexing schemes. Where funding to augment subject analysis is scarce, this kind of research-based information will assist in assessing whether free-text pre-iconographic description might be a more effective (and cost-effective) method of providing subject access than careful selected controlled vocabulary.

Explaining Search and Retrieval Mechanisms to Users

As we continue to grow and develop systems, it is critical that we share more information with users about how indexing is implemented in those systems. Howard Besser's (1997a) investigation of the seven implementations of the MESL project, by seven different institutions, using seven different indexing/search systems, underscores this point. This observation is validated by Shneiderman (1997), who states that in many systems there is little or no indication of how the system interprets a search request, so users have a difficult time interpreting search results.

Using Other Retrieval Mechanisms to Compensate for Lack of Semantic Indexing

In addition to conducting more empirical research into the effectiveness of semantic indexing, it would be useful to work more directly with computer scientists to evaluate the effectiveness of emerging computer-based retrieval mechanisms. This is an area of extremely active research in both universities and the commercial sector. During the MESL project, at least two research units at participating sites—Columbia University and the University of Illinois—did some experimentation with content-based retrieval that included the project's images. The potential for this kind of retrieval can be seen in Columbia's trio of projects—VisualSeek, WebSeek, and MetaSeek—which employ a variety of techniques for visual information retrieval, including incorporating user examples as input and matching them according to features such as color and texture (VisualSeek), combining text and color histogram searching (WebSeek), and using both visual content and keywords to search remote image collections with their own search engines (MetaSeek) (Benitez et al., 1998). There are few instances where research has been undertaken solely with images from cultural heritage collections.

Of even more potential impact on retrieval success, a number of

hybrid visual and semantic retrieval systems have been proposed or developed. Enser (1995) sets forth a conceptual model consisting of what he calls linguistic and visual search and query modes and explains how they might interact to improve search results. In such a system, a user could submit a search, select relevant images, and resubmit them to find related images based on the keyword associated with the visual surrogates (Mostafa, 1994). Other systems under development combine traditional text-based retrieval with elements of content-based retrieval; one such system is called SEMCOG or SEMantics and COGnition-based image retrieval (Li et al., 1997). In this model, a user may pose a query by combining textual descriptors, image content, and spatial relationships between objects. A similar experimental system has been built recently that draws on a test data set of cultural heritage information. The tool, called ARThur, was developed by the Getty Information Institute in cooperation with NEC using their Amore content-based retrieval system (http://www.isi.edu/cct/arthur/). ARThur allows a user to search by image content, contextual similarity (proximity between selected image and text on Web page), as well as by keyword. Keywords can be used to qualify queries by content to improve retrieval precision. The keyword searching is enhanced by allowing users to formulate queries using the Getty vocabulary tools, the *Union List of Artist Names* (ULAN), the *Getty Thesaurus of Geographic Names* (Getty TGN), and the *Art & Architecture Thesaurus* (AAT).

Capitalizing on Human Perceptual Capabilities

Writing in 1995, Donna Romer stated:

> Consistent and psychologically informed search models for multimedia retrieval are neither readily available nor obvious. The search models found in both early products and the research literature appear to be driven by what technology is able to do, rather than how people make perceptual sense of different modalities (pp. 50-51). . . . We have been proceeding into the multimedia age assuming that people "read" and understand images in the same way that they "read" and understand documents. (p. 50)

Current constraints on screen size and resolution limit the number of images that can be displayed at once; users are forced to page through screen after screen of thumbnails. The potential of the human eye/brain to make rapid relevancy judgments on images without reference to text needs to be exploited. Mechanisms to support "I'll know it when I see it" behavior could be developed to allow users to browse through large numbers of images quickly. The exploratory work done by the Maryland team with the National Digital Library (Marchionini, 1998), for instance, allows users to select viewing options to display up to fifty thumbnails at a time.

Supporting Functionality—Interface Design and Tool Development

Beyond the workings of discovery and retrieval, effective user interface design is critical in facilitating the use of the visual information returned by our image delivery systems. Returning to Sklar's (1995) functionality characteristics, she asserts that it is necessary to gain direct access to digital surrogates and be able to display a number of images on screen at once, allowing rapid and easy comparison (p. 14). If that functionality is characteristic of a "generic" interface, what other kinds of features of specialized interfaces might need to be developed to support specific users and uses?

Generic Interface Design Issues

Most of the Web-based image databases developed to date, including those developed in the MESL project, provide a series of common interfaces for returning search results to users. The Web browser serves as the basic interface with several more or less standard ways of displaying search results within the browser window. For browsing, a grid of thumbnails with brief identifying captions below is the most typical presentation; users must page through multiple screens if many thumbnails are returned in response to a query. For a more complete display, screen elements include fielded textual data presented in something like standard bibliographic format (brief display) with the associated image or images next to it. Options for viewing larger images and more textual information (full record) are often provided as well. Because many of these delivery systems have been designed in libraries and visual resource collections, it is not surprising that the general design grows directly from the library catalog tradition.

Lansdale et al. (1996) characterize this kind of design as "craft design," that is, evolutionary, developed by trial and error, where successful elements are incorporated and carried forward and unsuccessful ones drop away. They distinguish it from design grounded in scientific theoretical knowledge. Plaisant et al. (1995) note that designing image browsers involves many choices and requires more controlled experiments, prototyping, and validation. Mostafa (1994) also remarks on the rarity of theoretical and empirical research on user interface design in image retrieval systems. He urges additional exploration of our ability to process visual information rapidly, processing it only as visual information without reference to verbal descriptions (p. 118). More cooperation with human-computer interaction specialists as well as empirical research and usability testing will help validate effective design decisions and generate new models to be tested.

Functionality

In addition to displaying images in response to user queries, we need

to provide certain very basic tools to accompany the generic interface. For instance, the user needs to be able to select or "mark" records or images and to save them for later arrangement and manipulation. Providing the user with the ability to customize or choose display options is also useful. In the student and instructor evaluation of the MESL project, both instructors and students indicated "zooming in and out" and "comparing two images" as the most desirable display and manipulation features they wanted in future image databases. Instructors also expressed an interest in tools which would allow them to "save search results" and "sort and mark sets of images" (Sandore & Shaik, 1997).

A few of these functional tools are available in recently developed delivery systems. The AMICO interface, developed by RLG, includes the shopping cart function—called a "notebook" in the RLG system—where records and images can be saved during a session and then printed or downloaded. In addition, a menu selection called "Options" takes the user to a screen where they can select default viewing options for a session, including maximum image dimensions, sort order, number of items in each search result screen, and various other display options (see Figure 7). It is critical that the effectiveness of these kinds of basic tools be evaluated as well as determining whether others would add significantly to the functionality of our most basic interfaces. For instance, one could imagine the usefulness of extending user controls over interface options. Different interfaces might be selected depending on the information need (known item search versus browsing) or based on the number of hits in response to a query. Or an intelligent interface could be built that could respond to user input by selecting the most appropriate displays based on the user's path through the material.

Special Tools for Specific Users and Uses

In addition to investigating the effectiveness of search forms, browsing interfaces, and search result displays for generic applications such as union catalogs, further research is needed into the kinds of functionality, interfaces, and tools specific user groups need to enable specific uses of image databases.

The development of the ISIS software at Maryland in the MESL project clearly demonstrated that putting appropriate tools in the hands of faculty users greatly enhanced their ability to use both the visual information and the textual information in the database. This software successfully translated the process of selecting and arranging slides on a light table and placing them into slide carousels for classroom projection onto the computer screen (see Figure 8). Faculty members were presented with a familiar metaphor for their analog working environment and were therefore much more willing and able to utilize the underlying information in the data set.

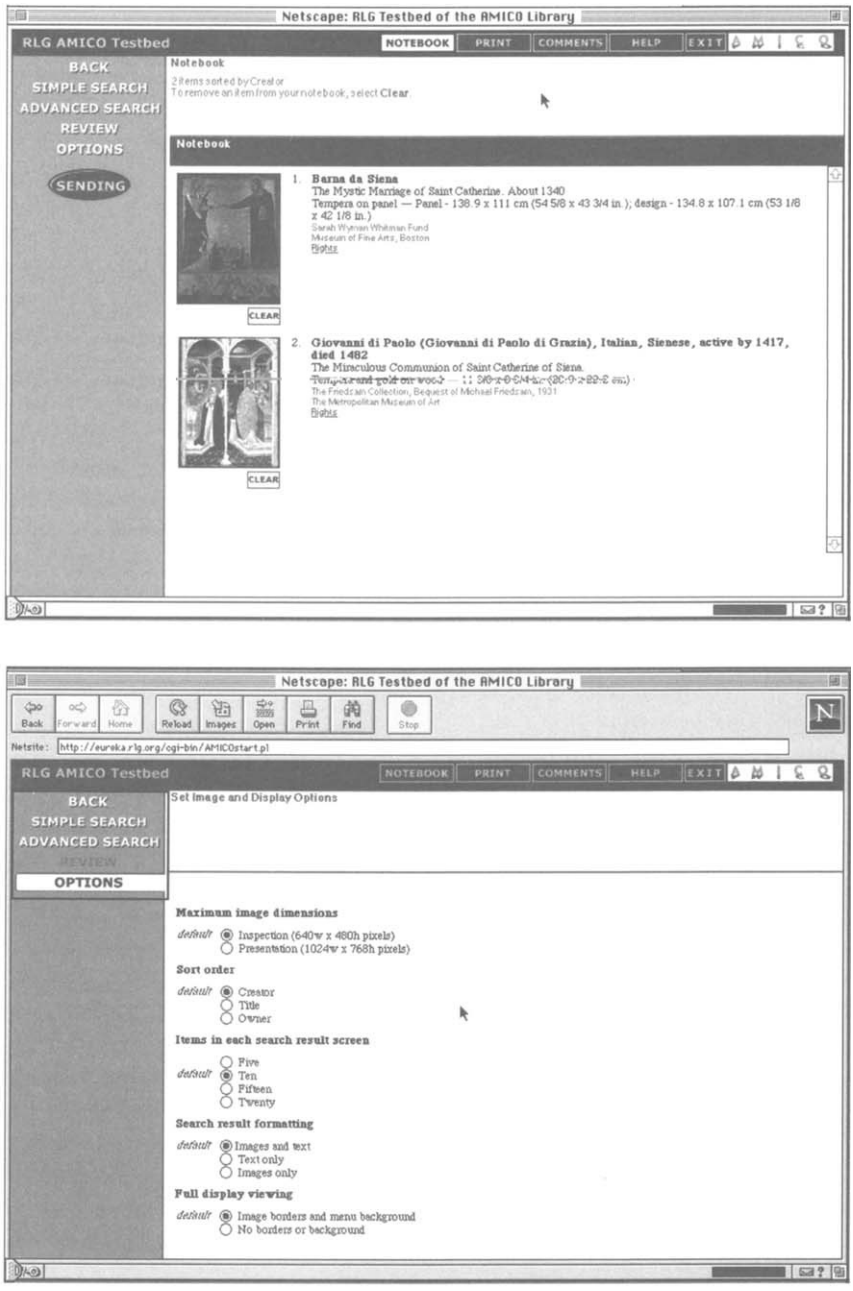


Figure 7. Screen Shots from RLG/AMICO Showing Notebook and Viewing Options.

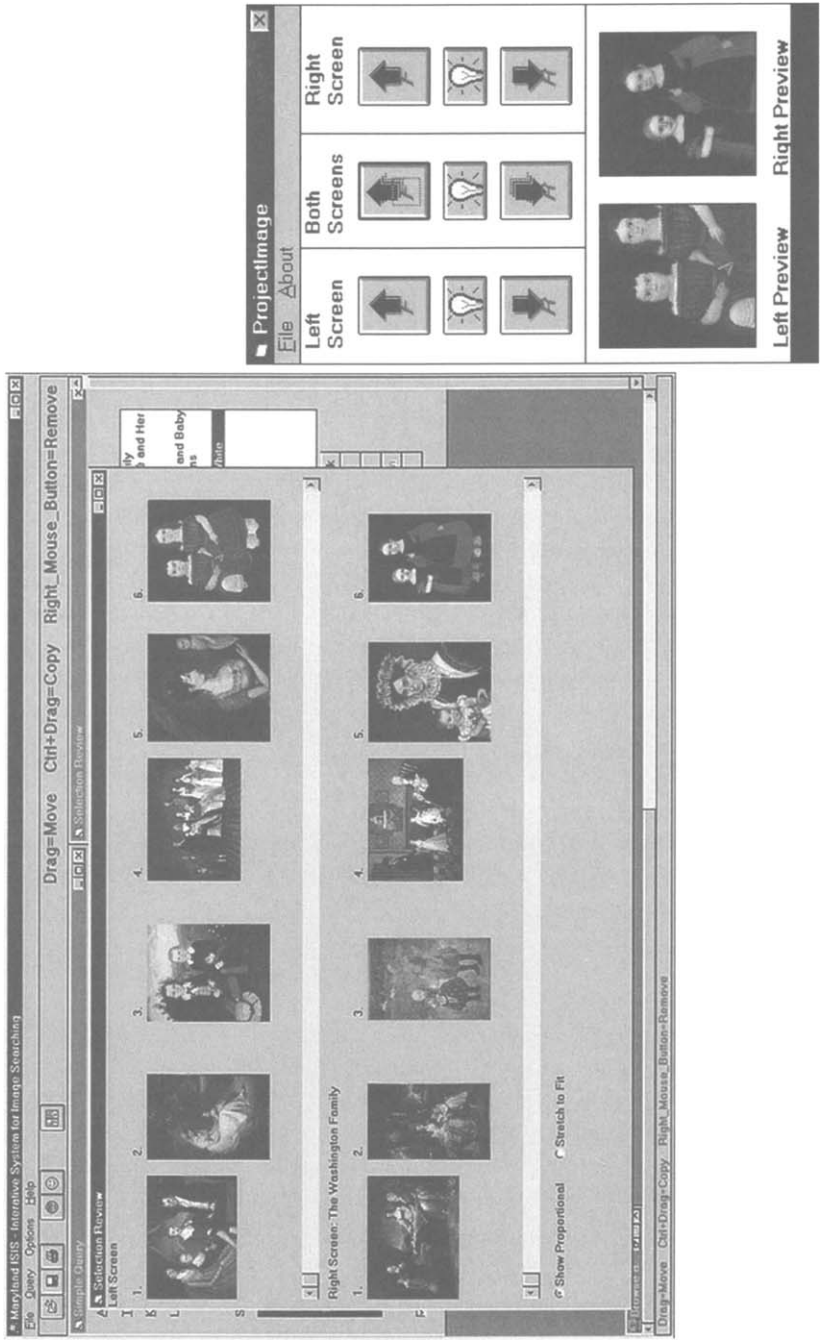


Figure 8. Maryland ISIS: Tools for Ordering Image Presentation in the Classroom and Controlling Projection.

For art historical research, other sets of tools would be required. Some of the modeling for these tools was undertaken in a joint project of the Getty Art History Information Program and Brown University's Institute for Research and Information (Bakewell et al., 1988). In a two-pronged study, the researchers investigated what art historians say they do in their research and observed what the art historians actually do in order to better understand what kinds of automated tools would enhance their work. The interview subjects reported a number of behaviors around which functionality could be built. They reported frequently writing in the margins of photocopies of works of art; collecting and arranging information by topic, not by format, commingling clippings, photographs, sales catalogs, and letters. They reported on building personal collections of images and frequently interfiling them with notes for a specific project. Rhyne (1998), in reviewing various image databases and museum sites on the Web, enumerates a list of basic requirements a scholar would expect from such sites. Although somewhat different in focus, Rhyne's article is an important example of the way specialist users can contribute to the articulation of system requirements.

For art historians, some of the behaviors that would need to be supported would be comparison, annotation, and the ability to examine details. Some of this functionality has begun to emerge in the past few years. Both the RLG/AMICO system and the Michigan Image Server give users instructions or tools to do side-by-side image comparison (see Figure 9).

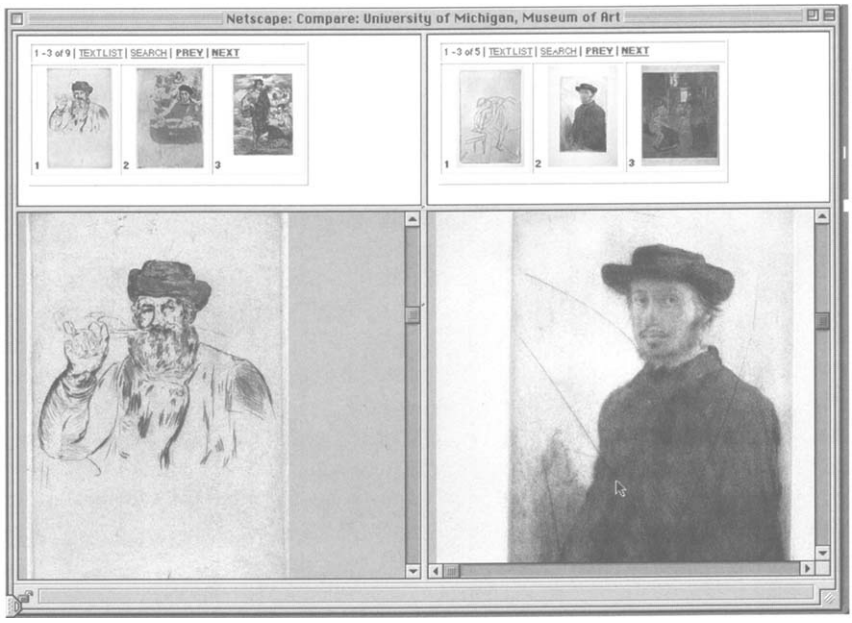


Figure 9. Michigan Image Services, Side by Side Comparison.

A JAVA-based image annotation tool, called I-NOTE, has been developed by the Institute for Advanced Technology in the Humanities at the University of Virginia (University of Virginia, 1998) (see Figure 10). Michigan is using Mr. SID, the wavelet compression software developed by LizardTech, to enable panning and zooming to examine image details.

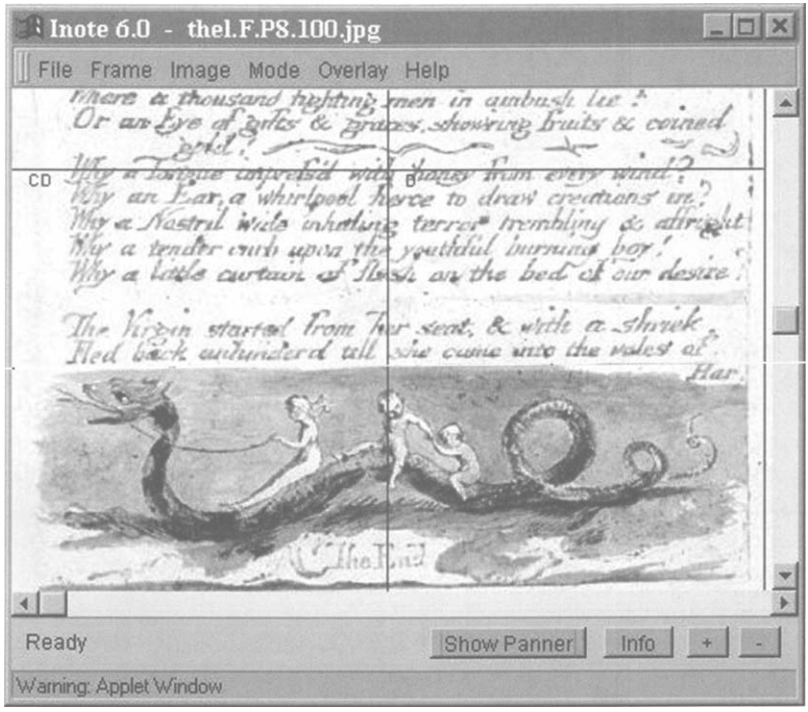


Figure 10. University of Virginia, Institute for Advanced Technology in the Humanities (IATH): I-Note, Image Annotation Tool, Screen Shot.

In the commercial sector, Luna Imaging, a digital imaging service provider founded by Michael Ester (formerly director of the Getty Art History Information Program), has developed an image management product called *Insight* aimed at the cultural heritage community. Presumably the design attempts to support some of the functionality identified in the AHIP/IRIS study. Luna's product literature states that *Insight* is designed with the image user, not the collection manager, in mind. It allows users to "look through materials, examine and compare images, view details, organize images into groups around ideas or events, and save subsets of images for particular applications." While products such as *Insight* may begin to meet users' functional requirements, it is important that as a

community we acknowledge the need to develop nonproprietary tool sets which can be put in the hands of all end-users.

As these functional tools continue to develop, we can envision the creation of customized tool sets appropriate for particular uses like personal research, courseware creation, organizing exhibitions, and the like. This is not simply a technical challenge; it is imperative that developers work closely with end-users, modeling and understanding how they work in the analog world and then engaging in an iterative development process in the digital environment. We need to acknowledge that "information systems and indexing tools designed for specific disciplines need to fit the needs of those fields rather than the 'typical' humanist scholar" (Tibbo, 1994, p. 608).

The process undertaken in the Getty AHIP/Brown IRIS project to understand the behaviors of art historians in their research needs to be extended to other disciplines where scholars make intensive use of visual materials so that those behaviors can be effectively supported by specialized interfaces and tool sets. As Ester and Shipp observed in their foreword to the AHIP/IRIS study: "Considering the magnitude of the commitments that institutions are making to automation, we would be well advised to improve our comprehension of the constituency we intend to serve" (Bakewell et al., 1988, p. xi). It seems likely that broad based transformations of scholarly research and teaching will be possible only after considerably more functionality is built into digital delivery systems.

Conclusion

The past few years have seen rapid growth in the development of image delivery systems to provide access to the wealth of cultural heritage information. The Museum Educational Site Licensing Project served as an important testbed, and successor projects are building on its successes and incorporating new functionality in their own systems. As the process of iterative development continues, it is critical that developers focus on users and uses and be explicit about the assumptions, biases, and limitations of their processes and their systems. In so doing, users will be better served and the documentation of system goals and design decisions will better inform future system development. In addition, more collaboration between information professionals, computer scientists, human-computer interaction specialists, instructional designers, and end-users (students, teachers, and scholars) will enrich and accelerate the development process. Those collaborations must include more quantitative and qualitative research, widely disseminated, in order to focus limited resources on those development efforts which will have the greatest impact on user success.

ACKNOWLEDGMENTS

The author would like to acknowledge the colleagues who provided suggestions, insights, support and editorial assistance: Howard Besser, Steve Chapman, Jennifer Vinopal, Beth Sandore, Don Waters, and John Weise. In addition, thanks to all my colleagues in the MESL project who provided a rich environment for exploration of the ideas developed in this discussion.

REFERENCES

Armitage, L. H., & Enser, P. G. B. (1997). Analysis of user need in image archives. *Journal of Information Science, 23*(4), 287-299.

Bakewell, E.; Beeman, W. O.; Reese, C. M.;, & Schmitt, M. (1988). *Object, image, inquiry: The art historian at work.* Santa Monica, CA: Getty Art History Information Program.

Benitez, A. B.; Beigi, M.; & Chang, S. F. (1998). Using relevance feedback in content-based image metasearch. *IEEE Internet Computing, 2*(4), 261-283.

Besser, H. (1997a). Comparing five implementations of the Museum Educational Site Licensing Project: If the museum's data's the same, why's it look so different? In *Proceedings of the Fourth International Conference on Hypermedia and Interactivity in Museums* (pp. 317-325). Pittsburgh, PA: Archives and Museum Informatics.

Besser, H. (1997b). Image databases: The first decade, the present, and the future. In P. B. Heidorn & B. Sandore (Eds.), *Digital image access and retrieval* (Proceedings of the 33rd Annual Clinic on Library Applications of Data Processing, March 24-26, 1996, University of Illinois at Urbana-Champaign). Urbana-Champaign: University of Illinois, Graduate School of Library and Information Science.

Besser, H. (1998). MESL implementation at the universities. In C. Stephenson & P. McClung (Eds.), *Delivering digital images: Cultural heritage resources for education* (pp. 70-84). Los Angeles: Getty Information Institute.

Borkowski, E. Y., & Hays, C. (1998). The Maryland Interactive System for Image Searching: Implementing a system for teaching with digital images. In P. McClung & C. Stephenson (Eds.), *Images online: Perspectives on the Museum Educational Site Licensing Project* (pp. 35-45). Los Angeles: Getty Information Institute.

Cawkell, A. E. (1993). Picture-queries and picture databases. *Journal of Information Science, 19,* 409-423.

Collins, K. (1998). Providing subject access to images: A study of user queries. *American Archivist, 61*(Spring), 36-55.

Dowden, R. (1998). The MESL data dictionary and the data export process. In C. Stephenson & P. McClung (Eds.), *Delivering digital images: Cultural heritage resources for education* (pp. 50-55). Los Angeles: Getty Information Institute.

Enser, P. G. B. (1995). Pictorial information retrieval. *Journal of Documentation, 51*(2), 126-170.

Ester, M. (1990). Image quality and viewer perception. In P. Grant-Ryan (Ed.), *Digital image–digital cinema* (SIGGRAPH '90 art show catalog. Leonardo Supplemental Issue) (pp. 51-63). Oxford: Pergamon Press.

Ester, M. (1994). Digital images in the context of visual collections and scholarship. *Visual Resources, 10*(1), 11-24.

Hastings, S. K. (1995). Query categories in a study of intellectual access to digitized art images. In T. Kinney (Ed.), *ASIS '95* (Proceedings of the 58th annual meeting of the American Society for Information Science, October 9-12, 1995, Chicago, IL) (pp. 3-8). Medford, NJ: American Society for Information Science.

Janney, K., & Sledge, J. (1995). *A user model for CIMI Z39.50 application profile.* Retrieved November 4, 1999 from the World Wide Web: http://www.cimi.org/documents/Z3950_app_profile_0995.html.

Jorgensen, C. (1996). Indexing images: Testing an image description template. In P. Solomon (Ed.), *ASIS '96* (Proceedings of the 59th annual meeting of the American Society for Information Science, October 21-24, 1996, Baltimore, MD) (pp. 209-213). Medford, NJ: American Society for Information Science.

Jorgensen, C. (1998). Attributes of images in describing tasks. *Information Processing & Management, 34*(2-3), 161-174.

Lansdale, M. W.; Scrivener, S. A. R.; & Woodcock, A. (1996). Developing practice with theory in HCI: Applying models of spatial cognition for the design of pictorial databases. *International Journal of Human-Computer Studies, 44*(6), 777-799.

Layne, S. (1994). Some issues in the indexing of images. *Journal of the American Society for Information Science, 45*, 583-588.

Li, W.-S.; Candan, K. S.; Hirata, K.; & Nara, Y. (1997). SEMCOG: An object-based image retrieval system and its visual query interface. In J. Peckham (Ed.), *Proceedings of the ACM SIGMOD International Conference on Management Data* (SIGMOD 1997, May 13-15, 1997, Tucson, AZ) (pp. 521-524). New York: ACM Press.

Lindermeier, R., & Stein, C. (1991). User requirement analysis in the museum and art history field for advanced computer system design. *Computers and the History of Art, 1*(2), 39-53.

Marchionini, G.; Plaisant, C.; & Komlodi, A. (1998). Interfaces and tools for the Library of Congress National Digital Library Program. *Information Processing & Management 34*(4), 535-555.

McClung, P. (Ed.). (1995). *RLG Digital Image Access Project.* Mountain View, CA: Research Libraries Group.

McClung, P., & Stephenson, C. (Eds.). (1998). *Images online: Perspectives on the Museum Educational Site Licensing Project.* Los Angeles: Getty Information Institute.

Mostafa, J. (1994). Digital image representation and access. In *Annual Review of Information Science and Technology* (vol. 29, pp. 91-135). Medford, NJ: Learned Information.

Mostafa, J., & Dillon, A. (1996). Design and evaluation of a user interface supporting multiple image query models. In S. Hardin (Ed.), *Proceedings of the ASIS Annual Meeting* (vol. 33, pp. 52-57). Medford, NJ: Information Today.

Ornager, S. (1997). Image retrieval: Theoretical analysis and empirical user studies. In C. Schwartz (Ed.), *Proceedings of the ASIS Annual Meeting* (vol. 34, pp. 202-211). Medford, NJ: Information Today.

Plaisant, C.; Carr, D.; & Shneiderman, B. (1995). Image-browser taxonomy and guidelines for designers. *IEEE Software, 12*(2), 21-32.

Promey, S. M. (1998). Digital images in the art history classroom: Personal reflections. In P. McClung & C. Stephenson (Eds.), *Images online: Perspectives on the Museum Educational Site Licensing Project* (pp. 13-22). Los Angeles: Getty Information Institute.

Rasmussen, E. M. (1997). Indexing images. In *Annual review of information science and technology* (vol. 32, pp. 169-196). Medford, NJ: Learned Information.

Rhyne, C. S. (1996). Computer images for research, teaching, and publications in art history and related disciplines. *Visual Resources, 12*, 19-51.

Rhyne, C. S. (1998). Images as evidence in art history and related disciplines. *VRA Bulletin, 25*(1), 58-66.

Romer, D. (1995). Image and multimedia retrieval. In *Research agenda for networked cultural heritage* (pp. 49-56). Santa Monica, CA: Getty Information Institute.

Sandore, B., & Shaik, N. (1997). *The use of an art image database in the classroom: Instructor and student evaluation report* (Museum Educational Site Licensing Project). Unpublished report, University of Illinois Library, Digital Imaging Initiative Program.

Shneiderman, B. (1997). Designing information-abundant Web sites: Issues and recommendations. *International Journal of Human-Computer Studies, 47*(1), 5-29.

Sklar, H. (1995). Why make images available online: User perspectives. In P. McClung (Ed.), *RLG Digital Image Access Project* (pp. 11-18). Mountain View, CA: Research Libraries Group.

Stam, D. (1994). Pondering pixeled pictures: Research directions in the digital imaging of art objects. *Visual Resources, 10*(1), 25-39.

Stebley, L. (1998). Faculty perspectives on teaching with digital images. In H. Besser & B. Yamashita (Eds.), *The cost of digital image distribution: The social and economic implications of the production, distribution and usage of image data* (Report to the Andrew W. Mellon Foundation). Berkeley: University of California, School of Information Management and Systems.

Stephenson, C., & McClung, P. (Eds.). (1998). *Delivering digital images: Cultural heritage resources for education.* Los Angeles: Getty Information Institute.

Tibbo, H. R. (1994). Indexing for the humanities. *Journal of the American Society for Information Science, 45*(8), 607-619.

Evaluation of Image Retrieval Systems: Role of User Feedback

SAMANTHA K. HASTINGS

ABSTRACT

INTELLECTUAL ACCESS TO A GROWING NUMBER OF NETWORKED image repositories is but a small part of the much larger problem of intellectual access to new information formats. As more and more information becomes available in digital formats, it is imperative that we understand how people retrieve and use images. Several studies have investigated how users search for images, but there are few evaluation studies of image retrieval systems. Preliminary findings from research in progress indicate a need for improved browsing tools, image manipulation software, feedback mechanisms, and query analysis. Comparisons are made to previous research results from a study of intellectual access to digital art images. This discussion will focus on the problems of image retrieval identified in current research projects, report on an evaluation project in process, and propose a framework for evaluation studies of image retrieval systems that emphasizes the role of user feedback.

INTRODUCTION

Problems with the retrieval of images are complicated by a lack of knowledge of how people search for, and use, images. There is a proliferation of image databases available on servers connected to the Web. As the number of images available increases, the more difficult it becomes to find the image that meets a specific information need. In addition, many of the documents that are being converted into electronic formats contain

Samantha K. Hastings, School of Library and Information Science, University of North Texas, P. O. Box 311068, Denton, TX 76203-1068
LIBRARY TRENDS, Vol. 48, No. 2, Fall 1999, pp. 438-452

images. Traditional retrieval and indexing methods for providing access to large text databases do not offer adequate access to the images. Text retrieval research has a history of several thousand years. Retrieval research for images has been going on for approximately ten years, and we are just now beginning to examine the content of images instead of viewing them as black boxes described by textual descriptors.

The differences between text and images necessitate that research in retrieval techniques for images begin with an understanding of how people search for images, how images are indexed, how images are used, input from users, and what manipulations of the images are needed for specific tasks. When the focus is narrowed to digital art images, the problem is even more complex because there are queries of art that are not specific or dependent on content. The investigation of intellectual access to art images is a small piece of the retrieval problem, but the nature of how people search art images reflects the difficulty of the problem. This is not just an indexing problem; sophisticated technology does not solve it, and it seems that pattern-matching algorithms only seem to work with known item searches.

Background

The major problems with the retrieval of digital images may be divided into four main categories: technical, semantic, content, and relativity. Technical problems include load time and bandwidth, lack of standard formats, color match systems, the size of image files in general, compression losses, and resolution variables. Most of these technical issues are capable of being resolved (Lynch, 1991; Besser & Trant, 1995). If we assume that bandwidth will increase, compression algorithms will improve, color match systems will be standardized, and needed resolutions will become available, then these technical problems should not consume us in the investigation of intellectual access to digital art images.

Semantic or concept-based problems deal with image retrieval terminology. Controlled vocabularies and standards to enable uniform access are used for concept-based indexing and retrieval. Projects such as the *Art and Architecture Thesaurus,* ICONCLASS, *The Thesaurus for Graphic Materials* (*TGM*), The Consortium for the Computer Interchange of Museum Information (CIMI), The Art Museum Image Consortium (AMICO), and many European projects attempt to standardize the language and retrieval mechanisms used to search for images (Barnett & Petersen, 1989; Busch, 1992; Moen, 1998).

We know that terms contained in a user's query are important indicators for indexed retrieval of images (Enser, 1995; Armitage & Enser, 1997; Jorgensen, 1996). Natural language searching is also investigated in a hypermedia environment with information in text nodes connected to an image for generation of a descriptor for the image (Dunlop & Van Rijsbergen, 1993). However, it is clear that using text to index a nontextual

medium leaves much to be desired. Enser (1995) states "linguistic identifiers, in the form of indexing terms, titles and captions, attached to images within a collection offer little promise as an effective pictorial information retrieval procedure" (p. 156).

Content-based issues in the retrieval of images are the current focus of at least twenty research groups (Gupta & Jain, 1997). Early research by Rorvig (1990) suggests that users presented with an image do not require textual descriptions. Content research includes systems that automatically identify and extract one or more of the following image attributes: color, shape, texture, spatial similarity, and text contained in an image. For example, Lunin (1994) presents a solid case for the use of texture for automated retrieval of fabric designs. Gupta and Jain (1997) give a detailed discussion of the capabilities of content-based image retrieval systems. Gudivada and Raghavan (1995) provide an excellent overview of the capabilities of content-based image retrieval systems. Examples include Query by Image Content (QBIC), ART MUSEUM using Query by Visual Example (QVE), CORE, the Chabot project from UC Berkeley, Virage for multimedia management, and Photobook. In addition to the work being conducted with still images, Goodrum (1997) and Turner (1995) have both looked at automatic indexing for video and moving images. Recently, Turner (1998) used closed-captioning as a source for index terms in the retrieval of moving images.

Of course, there are problems with content-based retrieval systems. For example, a search using the AltaVista search engine (which uses Virage for image retrieval) and limited to photos with the search term "Homer," retrieves two busts of the Greek Homer, six photos of Homer Simpson, a photo of a Winslow Homer painting, and so on. Most interesting is that when you click on "visually similar images" under a photo of a bust of the Greek Homer, the returns include many curious and questionable images but no other bronze busts or images of Homer.

The last category of problems in the retrieval of digital images deals with relativity issues. Relativity includes problems surrounding the *aboutness* of an image. Queries that deal with thematic and iconographical concepts or ask "Why is?" are particularly difficult to address in automated image retrieval systems. Shatford (1986) clearly interprets Panofsky's theory of meaning. Shatford distinguishes Panofsky's factual and expressional meaning as determining what the picture is of and what it is about. She concludes that, at the iconographical level, an image "cannot be indexed with any degree of consistency" (p. 45).

There are a number of user-centered approaches focused on query analysis and image retrieval tasks presented by Enser (1995), Hastings (1995), Jorgensen (1996), and Keister (1994). More work on user needs and query types in content-based retrieval is needed. Armitage and Enser (1997) continue their work with an additional collection of user queries

and a suggested matrix for classification of the query terms based on Panofsky's categories.

Based on the Jorgensen finding that category use may depend on the task in which a user is involved, Fidel (1997) questions "should the design and evaluation of image databases be guided by the tasks involved in image retrieval?" (p. 186). Using Jorgensen's attribute classes, Fidel analyzed 100 actual requests from an agency with a large collection of stock photos similar to the one in Enser's study. Fidel refines the question to whether performance measurements should apply to all retrieval tasks or "does each task require its own measurement?" (p. 186). The summary of searching-behavior characteristics is presented in the categories of *data pole* and *object pole.* In the *data pole,* images provide information, and relevance criteria can often be determined ahead of time. In the *object pole,* images are objects, relevance criteria are invoked when viewing the images, and browsing the whole answer set is required. Fidel concludes that, for the image-retrieval tasks analyzed in the study, "precision and recall as used for text retrieval might not be adequate tests in image retrieval" (p. 198).

O'Connor (in press) focuses on the users and uses to circumvent some of the difficulties in describing images in words. User generation of captions and verbal responses are gathered from a collection of 300 diverse images. The role of user feedback is highlighted in the belief that indexing must have an active functional quality to be effective (O'Connor, 1994). In addition, O'Connor is investigating the ability of people to rapidly browse many images without the constraints of categorizations. In this "show-me-the-pictures" approach lies great promise for increased retrieval effectiveness. Combined with user-supplied functional captions and responses, some of the problems and challenges inherent in the relativity category of image retrieval may be met.

However, the major problem of intellectual access to digitized images in a networked environment remains largely unsolved (Mostafa, 1994; Rasmussen, 1997). Reliable measures for evaluating image retrieval systems need to be developed or revised from text retrieval methods. We do know that providing surrogate or thumbnail representations of an image for browsing greatly improves access to a collection (Besser, 1990), but we are still unsure when and how to match the need to browse with the retrieval task or query.

Cawkell (1992) points out that co-citation patterns reveal very little communication and collaboration between the content-based and concept-based researchers. Unfortunately, this remains a difficult obstacle in the design and testing of image retrieval systems.

Current Study

In a previous study of intellectual access to digital art images, all aspects of search and retrieval in an art image database were analyzed (Hastings,

1994). The study investigated how variations in the retrieval parameters and access points affected the queries by art historians when they conduct research using an art image database. Access points include existing information about the collection such as artist, title, provenance, and suggestions from participants for additional access points. Categories of query complexity were compared to image complexities. The current study compares the findings from identified user queries, user-supplied access terms, and retrieval tasks on the Web to previous findings.

For the purposes of the current study, "intellectual access" is defined as the image searcher's ability to find and use (retrieve) the image that meets a stated need. A "query" consists of either a stated need or an expression of intended use. "Image" is used to represent a surrogate representation of a real painting. The following research questions frame the study:

1. Are there categories of queries that can be met by thumbnail (small surrogate) images?
2. Is there a relationship between queries and manipulation of images?
3. Do queries contain indicators to access points used for the retrieval of images?
4. Are there identifiable categories of images that increase the ability to browse a collection of images?
5. Are there identifiable image manipulations that need to be added to satisfy queries in the networked database of images?

Participants

The population of this study is image searchers on the Web. The subset of the population for this study is students in the School of Library and Information Science and the School of Visual Arts at the University of North Texas and members of the Image-L listserv. The selection of the sample within this population subset is based on subject interest (Caribbean paintings) and willingness of the subjects to participate in the study. It must be noted that the sample is self-selected, and sometimes it is not possible to match online survey data with interview data.

The Collection

The images used for this study are of paintings in the Bryant West Indies Collection housed in the Special Collections Department at the Main Library, University of Central Florida. There are sixty-six Caribbean paintings with a special focus on Haitian art. The collection contains paintings acquired from 1965 through 1990. Images of the paintings are stored on a Kodak Photo CD and are the property of the researcher. The images and thumbnails of the paintings are available in JPEG format at the University of North Texas Web site (http://www.unt.edu/Bryantart).

Procedures

The summer 1997 indexing and abstracting class at the School of Library and Information Science constructed a database of index fields for the digital images of the Bryant Collection of Caribbean Art. Each image record contains a unique image identifier (code), a corresponding thumbnail, and information for each index field. The fields include artist name, working title, index terms, abstracts, dimensions, and assigned categories for content and style. The user can view a high-resolution image of the painting by clicking on the thumbnail from the database template. Thumbnail images are available for browsing by random order (see Figure 1) and by categories of content or style. The project team assigned the categories of content and style.

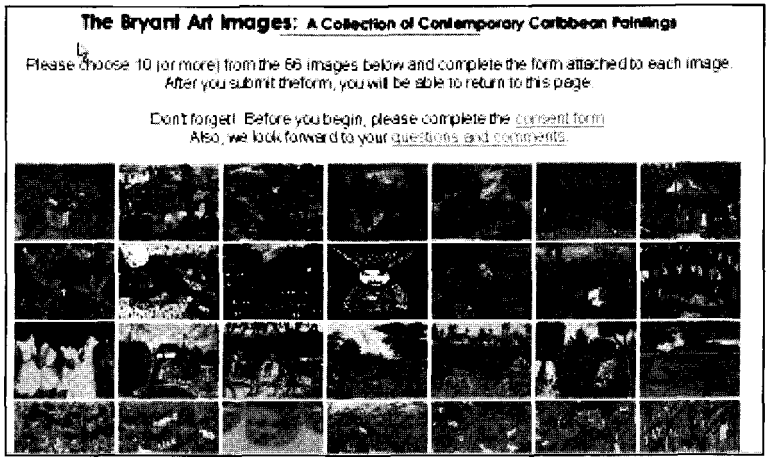


Figure 1. Thumbnail Images.

The index is assembled from controlled vocabularies and terms applied by the project team. The index includes thesaurus terms and is hypertext-linked to the thumbnail templates. In order to collect user-defined terms, a note form is included on each thumbnail template for searchers to add their own terms (see Figure 2). In addition, users are asked to rate the assigned index terms.

A user survey is available online and responses are sent to an e-mail account. T. J. Russell, research assistant, designed the Web pages. Russell conducted all pilot tests and contributed an integral part to the project. The introductory page for the project is represented in Figure 3. Survey and user-supplied data from approximately 200 responses are used for

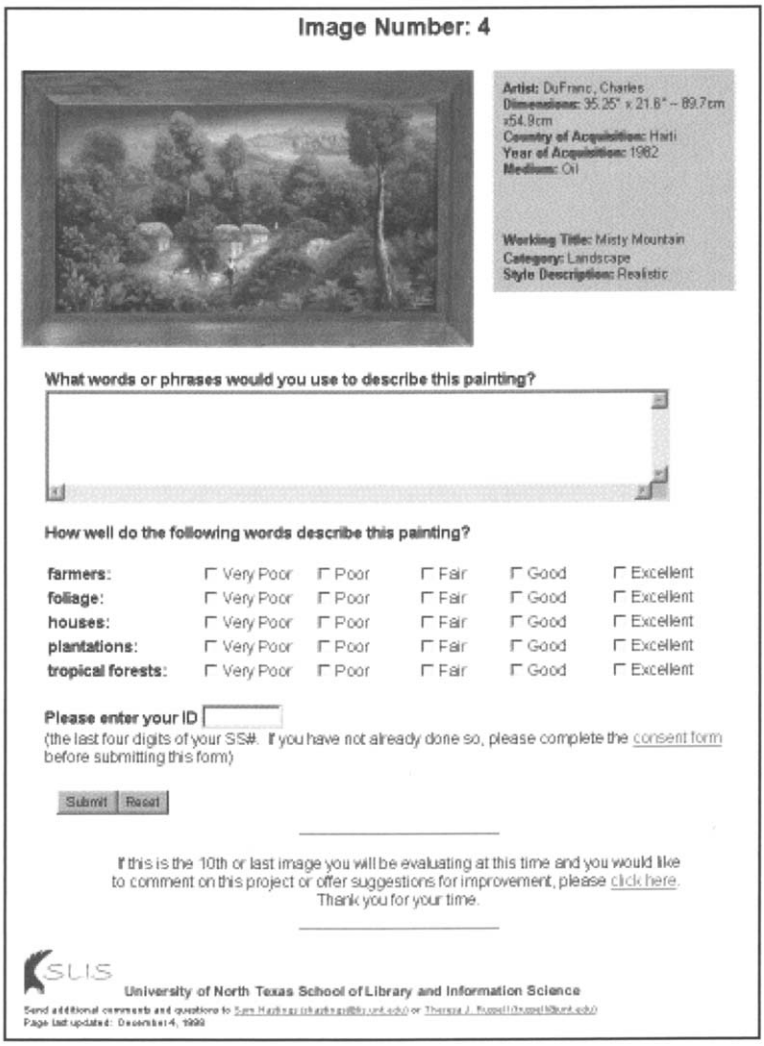
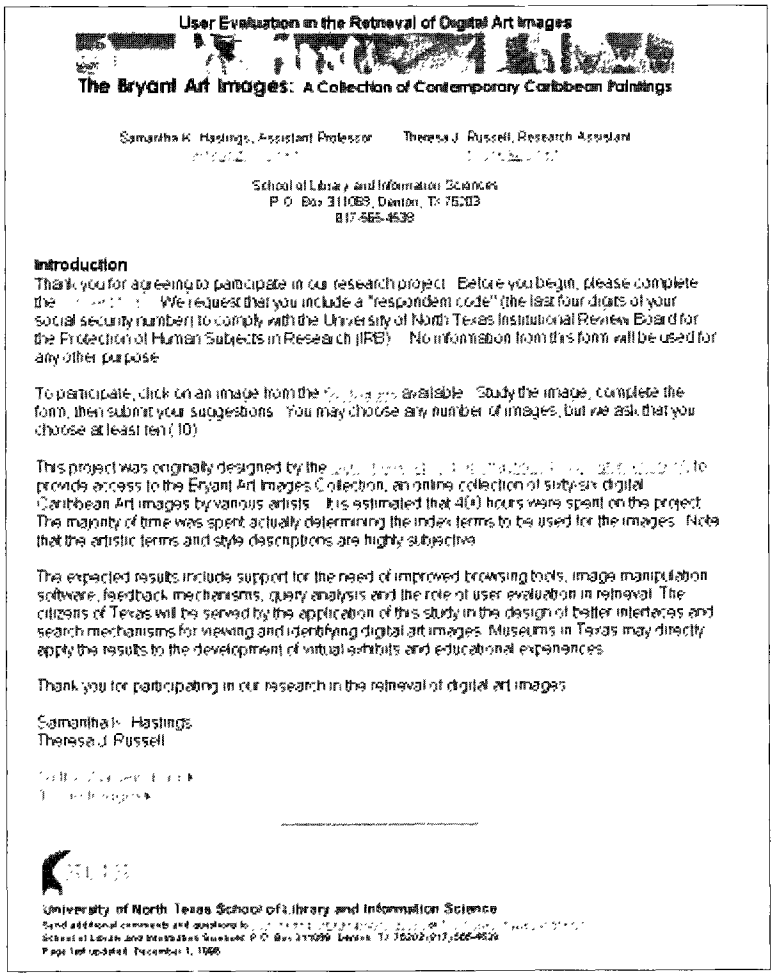


Figure 2. User Feedback Form and Survey.



User Evaluation in the Retrieval of Digital Art Images

The Bryant Art Images: A Collection of Contemporary Caribbean Paintings

Samantha K. Hastings, Assistant Professor Theresa J. Russell, Research Assistant

School of Library and Information Sciences
P. O. Box 311068, Denton, TX 76203
817-565-4539

Introduction

Thank you for agreeing to participate in our research project. Before you begin, please complete the [illegible]. We request that you include a "respondent code" (the last four digits of your social security number) to comply with the University of North Texas Institutional Review Board for the Protection of Human Subjects in Research (IRB). No information from this form will be used for any other purpose.

To participate, click on an image from the [illegible] available. Study the image, complete the form, then submit your suggestions. You may choose any number of images, but we ask that you choose at least ten (10).

This project was originally designed by the [illegible] to provide access to the Bryant Art Images Collection, an online collection of sixty-six digital Caribbean Art images by various artists. It is estimated that 40 hours were spent on the project. The majority of time was spent actually determining the index terms to be used for the images. Note that the artistic terms and style descriptions are highly subjective.

The expected results include support for the need of improved browsing tools, image manipulation software, feedback mechanisms, query analysis and the role of user evaluation in retrieval. The citizens of Texas will be served by the application of this study in the design of better interfaces and search mechanisms for viewing and identifying digital art images. Museums in Texas may directly apply the results to the development of virtual exhibits and educational experiences.

Thank you for participating in our research in the retrieval of digital art images.

Samantha K. Hastings
Theresa J. Russell

[illegible]

[illegible]

University of North Texas School of Library and Information Science

Send additional comments and questions to [illegible]
School of Library and Information Sciences, P. O. Box 311068, Denton, TX 76203-0917, 565-4539
Page last updated: December 1, 1998

Figure 3. Introductory Page.

the preliminary analysis reported below. Additional data are currently being collected. Analysis is an ongoing process, and the preliminary results reported here will be expanded.

Data Analysis

The data are being analyzed in three stages. First, the preliminary data from the online surveys and query statements are categorized and classified. The data are arranged in tables by query type. When possible, interview data are matched to each query, access points suggested, and image(s) used.

The second stage of analysis ranks user responses to existing index terms and looks for patterns in the searches for images on the Web. These patterns are derived from the tables produced in the first stage of data analysis. Relationships are noted for associations between query type and (1) display of the images; (2) access points or combinations of access points; and (3) stated requests for manipulations. The data are examined for patterns of variation.

The third stage of analysis compares the current data to previously collected data from a study of intellectual access to digital art images. Assertions were discovered from the analysis of the data and concepts were formed. The following concepts listed in Table 1 were developed from the assertions to describe the process of searching and retrieving digitized art images:

1. There are types and levels of queries used by art historians for searching photographic and digital art images.
2. The queries of art historians change when searching digital images. They become more complex, and they build on retrieved answer sets to create new queries.
3. There are computer functions needed for different levels of queries.
4. There is a relationship among level of query, access points, and computer manipulations for intellectual access to art images.
5. Some level one queries (see Table 1) can be answered without images.
6. Some level four queries (see Table 1) cannot be answered by the image or with primary textual information. Secondary subject resources are needed.
7. Digital images provide browse-style searchers with more opportunity to winnow for relevant retrieval sets.
8. Images can be described by level of complexity based on the analysis of color, composition, complexity, contrast, perspective, proportion, and style.
9. Queries of style retrieved more complex images.

Table 1 lists the major components of intellectual access identified in the

Table 1
MAJOR COMPONENTS OF INTELLECTUAL ACCESS TO DIGITAL ART IMAGES

Levels of Complexity	*Queries*	*Access Points*	*Computer Manipulations*
Level 1: Least Complex	Includes identification queries for who, where, when	Includes text fields and image in general	Use of search, sort, and display
Level 2: Complex	For queries of the type What are?"—requires sorting of the text information in the answer set	Includes sorted text information and images	Use of search, select, sort, display, and enlarge
Level 3: More Complex	Includes queries of style, subject, how, and ID of objects or activities	Includes style, keywords, and complex images	Use of compare, enlarge, mark, resolution, and style
Level 4: Most Complex	Includes queries for meaning, subject, and why	Includes style and subject	Use of style & subject searches plus access to full-text secondary subject resources

analysis of the study data by level of query complexity. Level one represents the least complex query level and level four represents the most complex. The table explains how the discovered concepts depend on complexity of the query and are linked to access points, computer manipulations, and traits of the image.

The previously defined categories showed a direct correlation between type of query and index access points and between type of query and complexity of image (Hastings, 1995). The results of the comparison to current data collected from the Web are discussed in the following section.

PRELIMINARY FINDINGS

The major difference in the data collected on the Web compared to previous data is the lack of ability to manipulate the images to meet the stated need in the query. Query categories for the Web searches fit into two categories. The first category is a combination of levels 1 and 2 (see

Table 1) from the previous study. Almost 60 percent of the queries collected asked for identification of the artist, activities, or place. The remaining 40 percent of the queries asked something about the subject of the painting, especially if the painting included voodoo ritual symbologies. This may change as we continue to collect and analyze data.

We are not able to compare computer manipulations or access points used at this time. Queries requiring a manipulation of the image to provide the answer could not be answered because the ability to compare images in sets and zoom-in or enlarge sections of the paintings was not possible.

The original research questions used to frame the current study are listed below with the findings we can support at this time:

1. *Are there categories of queries that can be met by thumbnail (small surrogate) images?* Almost 60 percent of the queries collected were answered with the use of thumbnail images. In the next stage of analysis, we will look at whether browsing the thumbnails could have answered the queries.
2. *Is there a relationship between queries and manipulation of images?* Several queries requested that portions of each image in a retrieved set be enlarged and compared on the same screen. The requested manipulations of the images were not available in this first set.
3. *Do queries contain indicators to the access points used in retrieving needed images?* For the queries that used text search terms, most of them appeared to have used the index and thesaurus as a guide in the formulation of the query.
4. *Are there identifiable categories of images that increase the ability to browse a collection of images?* The majority of users in the current set of data used the browse by category option, but it is unknown if that was from curiosity about the categories or from a relationship between their queries and the available categories. We do know from the survey and interview data that users suggest their own categories for sorting images for browsing and seemed to prefer the random categorization of images.
5. *Are there identifiable image manipulations that need to be added to meet queries in the networked database of images?* User notes from the online survey and interviews indicate that users need to be able to compare images, form images into sets for comparison, and have the ability to zoom-in or enlarge sections of the images.

IMPLICATIONS

The purpose of this study is to investigate how people query and retrieve digital art images on the Web. The study provides new information about the retrieval of images in a distributed network environment. How-

ever, there are also several problem areas discovered in this attempt to collect data from the Web. The very nature of the Web complicates the attempt to study how people access and use images because it is difficult to correlate online survey data with interview data. It is also difficult to separate duplicate responses. The Web environment presses the issue of testing because it continues to develop without waiting for the results from scholarly inquiry. Despite the complexities and lack of control over the environment, we are able to present three findings based on the data analysis.

We now know that browsing, manipulation of the images, and need for user interaction are important aspects of the search for images on the Web. As discussed in the implications section above, the capabilities to zoom-in on, enlarge, and group the images were not available on the Web. Image searchers on the Web need the additional capabilities that such software offers. For example, users with queries about the style of a painting often want to zoom-in on, and enlarge, an area to study color or brush strokes. Queries from the "compare" category need to be able to group different sets of images for comparison. It is especially important for users to be able to move and manipulate high-resolution images, not just the thumbnails. The conclusion is that the more complex the query, the more options for manipulation are required.

The responses collected from the survey form indicate the need for users to add their own descriptors and index terms in the search process. The application of relevance feedback mechanisms needs to dramatically improve. As we continue to collect and analyze the terms supplied by the users of the Caribbean art images, we will look for patterns or relationships between the supplied terms and the query.

The ability to browse the images becomes even more important on the Web. Thumbnail surrogates, as representations of the high-resolution image, are used as access points. However, thumbnails as surrogates present their own problems. Automatic extractions often capture only part of the high-resolution image, and there is little control over what part is used. We need to look at the importance of thumbnail categories to aid browsing. So far, there are more users of the random browse category than the supplied categories of content and style. It is important that users have the capability of applying their own categories for sorting and browsing. It may be that there are indicators in a query that system designers can use to supply possible categories.

Finally, the whole problem of "relativity" or queries of "why" is largely unsolved. We are finding some attempts by users to add dimensions of their own knowledge to the subject of a painting—especially for queries about meaning in the paintings, such as voodoo rituals. It is this role for user feedback that brought on the discussion of what is needed to effectively evaluate an image retrieval system.

A Suggested Framework for Evaluation Studies

Based on the work of the researchers mentioned in the background section of this article and the preliminary results of the current study, a combination of methods for evaluation of image retrieval systems are suggested in Table 2.

Table 2.
Framework for Evaluation of Image Retrieval Systems

Query or Retrieval Task	*Retrieval or Search Tools*	*Evaluation Method*
Identification of known item or image	Index text and fields Browse images	User & relevance feedback Relevant? Yes or No Measures of time & effort
Identification of unknown item(s) in image and/or index	Select & display sets of images Sort sets Enlarge	User supplied terms & categories for browsing Survey form Online user feedback mechanisms Measures of time & effort
Investigations of style and image content	Content-based retrieval tools such as color, texture, shape, and so on	Log analysis Screen captures Survey form
Queries asking "why" and investigations for "aboutness"	Random browsing and extensive answer set displays	Amount of user effort Observation of browsing behavior and answer set development
	May require secondary resources—e.g., biographical and historical information	Capture retrieved sets and compare to query/task

The important questions that arise from the suggested framework are:

- How and when are user feedback mechanisms that include opportunities for user knowledge added to the database?
- What is the nature of browsing in an image database and what types of flexibility need to be inherent in the system?

- What types of manipulation of the images are needed and when?

and finally,

- How does user interaction and feedback improve the retrieval of images?

REFERENCES

Armitage, L. H., & Enser, P. G. B. (1997). Analysis of user need in image archives. *Journal of Information Science, 23*(4), 287-299.

Barnett, P. J., & Petersen, T. (1989). Subject analysis and AAT/MARC implementation. *Art Documentation, 8*(4), 171-190.

Besser, H. (1990). Visual access to visual images: The UC Berkeley image database project. *Library Trends, 38*(4), 787-798.

Besser, H., & Trant, J. (1995). *Introduction to imaging: Issues in constructing an image database.* Santa Monica, CA: The Getty Art History Information Program.

Busch, J. A. (1992). Overview of art information endeavors. *Bulletin of the American Society for Information Science, 18,* 8-13.

Cawkell, A. E. (1992). Selected aspects of image processing and management: Review and future prospects. *Journal of Information Science, 18*(3), 179-192.

Dunlop, M. D., & VanRijsbergen, C. J. (1993). Hypermedia and free text retrieval. *Information Processing & Management, 29*(3), 287-298.

Enser, P. G. B. (1995). Pictorial information retrieval. *Journal of Documentation, 51*(2), 126-170.

Fidel, R. (1997). The image retrieval task: Implications for the design and evaluation of image databases. *New Review of Hypermedia and Multimedia, 3,* 181-199.

Goodrum, A. (1997). *Evaluation of text-based and image-based representations for moving image documents.* Unpublished doctoral dissertation, University of North Texas, Denton.

Gudivada, V. N., & Raghavan, V. V. (1995). Content-based image retrieval systems. *Computer, 28*(9), 18-22.

Gupta, A.; Santini, S.; & Jain, R. (1997). In search of information in visual media. *Communications of the ACM, 40*(12), 34-42.

Hastings, S. K. (1994). *An exploratory study of intellectual access to digitized art images.* Unpublished doctoral dissertation, Florida State University, Tallahassee.

Hastings, S. K. (1995). Query categories in a study of intellectual access to digitized art images. In T. Kinney (Ed.), *ASIS '95* (Proceedings of the 58th annual meeting of the American Society for Information Science, October 9-12, 1995, Chicago, IL) (pp. 3-8). Medford, NJ: American Society for Information Science.

Jorgensen, C. (1996). Indexing images: Testing an image description template. In P. Solomon (Ed.), *ASIS '96* (Proceedings of the 59th annual meeting of the American Society for Information Science, October 21-24, 1996, Baltimore, MD) (pp. 209-213). Medford, NJ: American Society for Information Science.

Keister, L. H. (1994). User types and queries: Impact on image access systems. In R. Fidel, T. Bellardo Hahn, E. M. Rasmussen, & P. J. Smith (Eds.), *Challenges in indexing electronic text and images* (pp. 7-22). Medford, NJ: Learned Information.

Layne, S. S. (1986). Analyzing the subject of a picture: A theoretical approach. *Cataloging & Classification Quarterly, 6*(3), 39-62.

Lunin, L. (1994). Analyzing art objects for an image database. In R. Fidel, T. Bellardo Hahn, E. M. Rasmussen, & P. J. Smith (Eds.), *Challenges in indexing electronic text and images* (pp. 57-72). Medford, NJ: Learned Information.

Lynch, C. A. (1991). The technologies of electronic imaging. *Journal of the American Society for Information Science, 42*(8), 578-585.

Moen, W. E. (1998). Accessing distributed cultural heritage information. *Communications of the ACM, 41*(4), 44-48.

Mostafa, J. (1994). Digital image representation and access. *Annual Review of Information Science and Technology, 29,* 91-135.

Panofsky, E. (1955). *Meaning in the visual arts: Papers in and on art history.* Garden City, NY: Doubleday.

Rasmussen, E. M. (1997). Indexing images. *Annual Review of Information Science and Technology, 32*, 169-196.

Rorvig, M. E. (1990). Intellectual access to graphic information (issue theme). *Library Trends, 38*(4), 639-815.

Turner, J. (1995). Comparing user-assigned terms with indexer-assigned terms for storage and retrieval of moving images: Research results. In T. Kinney (Ed.), *ASIS '95* (Proceedings of the 58th annual meeting of the American Society for Information Science, October 9-12, 1995, Chicago, IL) (pp. 9-12). Medford, NJ: American Society for Information Science.

Additional References

Enser, P. G. B. (1993). Query analysis in a visual information retrieval context. *Journal of Document and Text Management, 1*(1), 25-52.

Layne, S. S. (1994). Some issues in the indexing of images. *Journal of the American Society for Information Science, 45*(8), 583-588.

Markey, K. (1986). *Subject access to visual resources collections: A model for computer construction of thematic catalogs.* New York: Greenwood Press.

O'Connor, B. C. (1996). *Explorations in indexing and abstracting: Pointing, virtue, and power.* Englewood, CO: Libraries Unlimited.

O'Connor, B. C.; O'Connor, M. K.; & Abbas, J. M. (1999). User reactions to access mechanisms: An exploration based on captions for images. *Journal of the American Society for Information Science, 50*(8), 681-697.

Experimental Approaches

"Information Retrieval Beyond the Text Document," *Yong Rui, Michael Ortega, Thomas S. Huang, and Sharad Mehrotra*

"Precise and Efficient Retrieval of Captioned Images: The MARIE Project," *Neil C. Rowe*

"Exploiting Multimodal Context in Image Retrieval," *Rohini K. Srihari and Zhongfei Zhang*

Information Retrieval Beyond the Text Document

YONG RUI, MICHAEL ORTEGA, THOMAS S. HUANG, AND SHARAD MEHROTRA

ABSTRACT

WITH THE EXPANSION OF THE INTERNET, searching for information goes beyond the boundary of physical libraries. Millions of documents of various media types—such as text, image, video, audio, graphics, and animation—are available around the world and linked by the Internet. Unfortunately, the state of the art of search engines for media types other than text lags far behind their text counterparts. To address this situation, we have developed the Multimedia Analysis and Retrieval System (MARS). This article reports some of the progress made over the years toward exploring information retrieval beyond the text domain. In particular, the following aspects of MARS are addressed in the article: visual feature extraction, retrieval models, query reformulation techniques, efficient execution speed performance, and user interface considerations. Extensive experimental results are reported to validate the proposed approaches.

INTRODUCTION

Huge amounts of digital data are being generated daily. Scanners convert the analog/physical data into digital form; digital cameras and camcorders directly generate digital data at the production phase. Owing to all these multimedia devices, presently information is in all media types, including graphics, images, audio, and video in addition to the conventional text media type. Not only is multimedia information being generated

Yong Rui, Microsoft Research, One Microsoft Way, Redmond, WA 98052
Michael Ortega, 444 Computer Science, University of California, Irvine, CA 92697-3425
Thomas S. Huang, Beckman Institute for Advanced Science and Technology, University of Illinois, Urbana, IL 61801
Sharad Mehrotra, Department of Information and Computer Science, University of California, Irvine, CA 92697-3425
LIBRARY TRENDS, Vol. 48, No. 2, Fall 1999, pp. 455-474

at an ever-increasing rate, it is transmitted worldwide due to the expansion of the Internet. Experts say that the Internet is the largest library that ever existed; it is, however, also the most disorganized library ever.

Textual document retrieval has achieved considerable progress over the past two decades. Unfortunately, the state of the art of search engines for media types other than text lags far behind their text counterparts. Textual indexing of nontextual media, although common practice, has some limitations. The most notable limitations include the human effort required and the difficulty of describing accurately certain properties humans take for granted while having access to the media. Consider how human indexers would describe the ripples on an ocean; these could be very different under situations such as calm weather or a hurricane. To address this situation, we undertook the Multimedia Analysis and Retrieval System (MARS) project to provide retrieval capabilities to rich multimedia data. Research in MARS addresses several levels including the multimedia features extracted, the retrieval models used, query reformulation techniques, efficient execution speed performance, and user interface considerations.

This article reports some of the progress made over the years toward exploring information retrieval (IR) beyond the text domain. In particular, the discussion will concentrate on visual information retrieval (VIR) concepts as opposed to implementation issues. MARS explores many different visual feature representations. A review of these features appears in the next section ("Visual Feature Extraction"). These visual features are analogous to keyword features in textual media. Another section ("Retrieval Models Used in MARS") describes two broad retrieval models we have explored: the Boolean and vector models and the incorporated enhancements to support visual media retrieval such as relevance feedback. Results are given in a later section ("Experimental Results"). The last section provides remarks summarizing the overall discussion ("Conclusion").

Visual Feature Extraction

The retrieval performance of any IR system is fundamentally limited by the quality of the "features" and the retrieval model it supports. This section sketches the features obtained from visual media. In text-based retrieval systems, features can be keywords, phrases, or structural elements. There are many techniques for reliably extracting, for example, keywords from text documents. The *visual counterparts* to textual features in visual based systems are features such as color, texture, and shape.

For each feature, there are several different techniques for representation. The reason for this is twofold: (1) the field is still under development and, more importantly, (2) features are perceived differently by people and thus different representations cater to various preferences. Image features are generally considered as orthogonal to each other. The

idea is that a feature will capture some dimension of the content of the image, and different features will effectively capture different aspects of the image content. In this way, two images closely related in one feature could be very different in another feature. A simple example of this are two images, one of a deep blue sky and the other of a blue ocean. These two images could be very similar in terms of just color; however, the ripples caused by waves in the ocean add a distinctive pattern that distinguishes the two images in terms of their texture. Rui et al. (1999) give a detailed description of the visual features, and the following paragraphs emphasize the important ones.

The *color* feature is one of the most widely used visual features in VIR. This feature captures the color content of images. It is relatively robust to background complication and independent of image size and orientation. Some representative studies of color perception and color spaces can be found in McCamy et al. (1976) and Miyahara (1988). In VIR, color histograms (Swain & Ballard, 1991), color moments (Stricker & Orengo, 1995), and color sets (Smith & Chang, 1995) are the most used representations.

Texture refers to the visual patterns that have properties of homogeneity that do not result from the presence of only a single color or intensity. It is an innate property of virtually all surfaces, including clouds, trees, bricks, hair, fabric, and so on. It contains important information about the structural arrangement of surfaces and their relationship to the surrounding environment (Haralick et al., 1973). Co-occurrence matrix (Haralick et al., 1973), Tamura texture (Tamura et al., 1978), and Wavelet texture (Kundu & Chen, 1992) are the most popular texture representations.

In general, the *shape* representations can be divided into two categories: boundary-based and region-based. The former uses only the outer boundary of the shape while the latter uses the entire shape region (Rui et al., 1996). The most successful representatives for these two categories are Fourier Descriptor and Moment Invariants. Some recent work in shape representation and matching includes the Finite Element Method (FEM) (Pentland et al., 1996), Turning Function (Arkin et al., 1991), and Wavelet Descriptor (Chuang & Kuo, 1996).

Retrieval Models Used in MARS

With the large number of retrieval models proposed in the IR literature, MARS attempts to exploit this research for content-based retrieval over images. The retrieval model comprises the document or object model (here a collection of feature representations), a set of feature similarity measures, and a query model.

The Object Model

We first need to formalize how an object is modeled (Rui et al., 1998b). We will use images as an example, even though this model can be used for

other media types as well. An image object O is represented as:

$$O = O(D, F, R) \quad (1)$$

- D is the raw image data—e.g., a jpeg image.
- $F = \{f_i\}$ is a set of low-level visual features associated with the image object, such as color, texture, and shape.
- $R = \{r_{ij}\}$ is a set of representations for a given feature f_i—e.g., both color histogram and color moments are representations for the color feature (Swain & Ballard, 1991).

  Note that, each representation r_{ij} itself may be a vector consisting of multiple components, that is:

$$r_{ij} = [r_{ij1}, \ldots r_{ijk}, \ldots r_{ijK}] \quad (2)$$

  where K is the length of the vector.

Figure 1 shows a graphic representation of the object (image) model. The proposed object model supports multiple representations to accommodate the rich content in the images. An image is thus represented as a collection of low-level image feature representations (see section entitled "Visual Feature Extraction") extracted automatically using computer vision methods as well as a manual text description of the image.

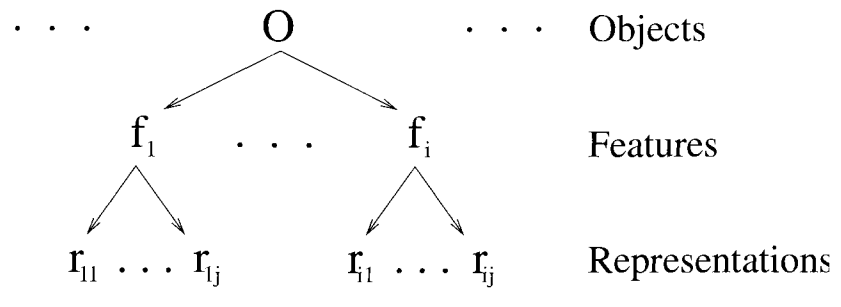


Figure 1. The Object Model.

Each feature representation is associated with some similarity measure. All these similarity measures are normalized to lie within [0,1] to denote the degree to which two images are similar in regard to the same feature representation. A value of 1 means that they are very similar and a value of 0 means that they are very dissimilar. Revisiting our blue sky and ocean example from the early section ("Visual Feature Extraction"), the sky and ocean images may have a similarity of 0.9 in the color histogram representation of color and 0.2 in the wavelet representation of texture. Thus the two images are fairly similar in their color content but very different in their texture content. This mapping $M = \{$<*feature representation,*

similarity measure$_i$>, ...} together with the object model O, forms (D, F, R, M), a foundation on which query models can be built.

Query Models

Based on the *object model* and the *similarity measures* defined above, query models that work with these raw features are built. These query models, together with the object model, form complete retrieval models used for VIR.

We explore two major models for querying. The first model is an adaptation of the Boolean retrieval model to visual retrieval in which selected features are used to build predicates used in a Boolean expression. The second model is a vector (weighted summation) model where all the features of the query object play a role in retrieval. The section on Boolean retrieval describes the Boolean model and the section on the "Vector Model" describes that model.

Boolean Retrieval

A user may not only be interested in more than a single feature from a single image. It is very likely that the user may choose multiple features from multiple images. For example, using a point-and-click interface, a user can specify a query to retrieve images similar to an image A in color and similar to an image B in texture. To cope with composite queries, a Boolean retrieval model is used to interpret the query and retrieve a set of images ranked based on their similarity to the selected feature.

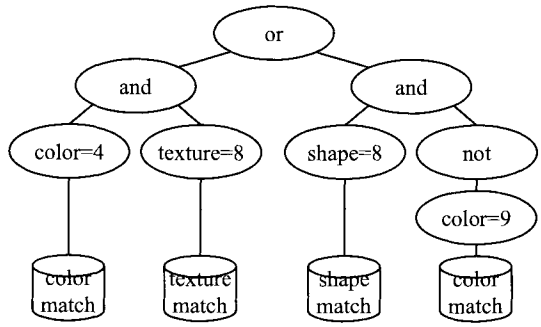
The basic Boolean retrieval model needs a pre-defined threshold, which has several potential problems (Ortega et al., 1998b). To overcome these problems, we have adopted the following two extensions to the basic Boolean model to produce a ranked list of answers:

- **Fuzzy Boolean Retrieval.** The similarity between the image and the query feature is interpreted as the degree of membership of the image to the fuzzy set of images that match the query feature. Fuzzy set theory is used to interpret the Boolean query, and the images are ranked based on their degree of membership in the set.
- **Probabilistic Boolean Retrieval.** The similarity between the image and the query feature is considered to be the probability that the image matches the user's information need. Feature independence is exploited to compute the probability of an image satisfying the query which is used to rank the images.

In the discussion below, we will use the following notations. Images in the collection are denoted by $I_1, I_2, \ldots I_m$. Features over the images are denoted by $F_1, F_2, \ldots F_r$, where F_i denotes both the name of the feature as well as the domain of values that the feature can take. The j^{th} instance of feature F_i corresponds to image I_j and is denoted by f_{ij}. For example, say F_1 is the color feature which is represented in the database using a histogram.

In that case, F_1 is also used to denote the set of all the color histograms, and $f_{1,5}$ is the color histogram for image 5. Query variables are denoted by $v_1, v_2, \ldots v_n \mid v_k \in F_i$ so each v_k refers to an instance of a feature F_i (an f_{ij}). Note that $F_i(I_j) = f_{ij}$. During query evaluation, each v_k is used to rank images in the collection based on the feature domain of f_i (F_i), that is v_k's domain. Thus, v_k can be thought of as being a list of images from the collection ranked based on the similarity of v_k to all instances of F_i. For example, say F_2 is the set of all wavelet texture vectors in the collection, if $v_k = f_{2,5}$, then v_k can be interpreted as being both the wavelet texture vector corresponding to image 5 and the ranked list of all $<I, S_{F2}(F_2(I), f_{2,5})>$ with S_{F2} being the similarity function that applies to two texture values.

A query $Q(v_1, v_2, \ldots v_n)$ is viewed as a query tree whose leaves correspond to single feature variable queries. Internal nodes of the tree correspond to the Boolean operators. Specifically, nonleaf nodes are one of three forms: $(v_1, v_2, \ldots v_n)$: a conjunction of positive literals; $(v_1, v_2, \ldots v_p, v_{p+1}, \ldots v_n)$, a conjunction consisting of both positive and negative literals; and $(v_1, v_2, \ldots v_n)$, which is a disjunction of positive literals. The following is an example of a Boolean query: $Q(v_1, v_2) = (v_1 = f_{1,5}) \wedge (v_2 = f_{2,6})$ is a query where v_1 has a value equal to the color histogram associated with image I_5, and v_2 has a value of the texture feature associated with I_6. Thus, the query Q represents the desire to retrieve images whose color matches that of image I_5 *and* whose texture matches that of image I_6. Figure 2 shows an example query $Q(v_1, v_2, v_3, v_4) = ((v_1 = f_{1,4}) \wedge (v_2 = f_{2,8})) \vee ((v_3 = f_{3,8}) \wedge \neg (v_4 = f_{1,9}))$ in its tree representation.



Operators: And, Or, Not
Basic features and representations:
Color histogram, color moment, wavelet texture, ...

Figure 2. Sample Query Tree.

Weighting in the Query Tree

In a query, one feature can receive more importance than another according to the user's perception. The user can assign the desired importance to any feature by a process known as *feature weighting*. Traditionally, retrieval systems (Flickner et al., 1995; Bach et al., 1996) use a linear scaling factor as feature weights. Under our Boolean model, this is not desirable. Fagin and Wimmers (1997) noted that such linear weights do not scale to arbitrary functions used to compute the combined similarity of an image. The reason is that the similarity computation for a node in a query tree may be based on operators other than a weighted summation of the similarity of the children. Fagin and Wimmers (1997) present a way to extend linear weighting to the different components for arbitrary scoring functions as long as they satisfy certain properties. We are unable to use their approach since their mapping does not preserve orthogonality properties on which our algorithms rely (Ortega et al., 1998b). Instead, we use a mapping function from $[0,1] \rightarrow [0,1]$ of the form:

$$similarity' = similarity^{\frac{1}{weight}}, 0 < weight < \infty \tag{3}$$

which preserves the range boundaries [0,1] and boosts or degrades the similarity in a smooth way. Sample mappings are shown in Figure 3. This method preserves most of the properties explained in Fagin and Wimmers (1997), except it is undefined for a weight of 0. In Fagin and Wimmers, a weight of 0 means the node can be dismissed. Here, $\lim_{weight \rightarrow 0} similarity' = 0$ for $similarity \in [0,1)$. A perfect similarity of 1 will remain at 1. This mapping is performed at each link connecting a child to a parent in the query tree.

Figure 4a shows how the fuzzy model would work with our running example of blue sky and blue ocean images. Figure 4b shows how the probabilistic model would work with our running example of blue sky and blue ocean images.

Computing Boolean Queries

Fagin (1996) proposed an algorithm to return the top k answers for queries with monotonic scoring functions that has been adopted by the Garlic multimedia information system under development at the IBM Almaden Research Center (Fagin & Wimmers, 1997). A function F is monotonic if $F(x_1, \ldots x_m) \leq F(x'_1, \ldots x'_m)$ for $x_i \leq x'_i$ for every i. Note that the scoring functions for both conjunctive and disjunctive queries for the fuzzy and probabilistic Boolean models satisfy the monotonicity property. This algorithm relies on reading a number of objects from each branch in the query tree until it has k objects in the intersection. Then it falls back on probing to enable a definite decision. In contrast, our algorithms (Ortega et al., 1998b) are tailored to specific functions that combine object scoring (here called fuzzy and probabilistic models).

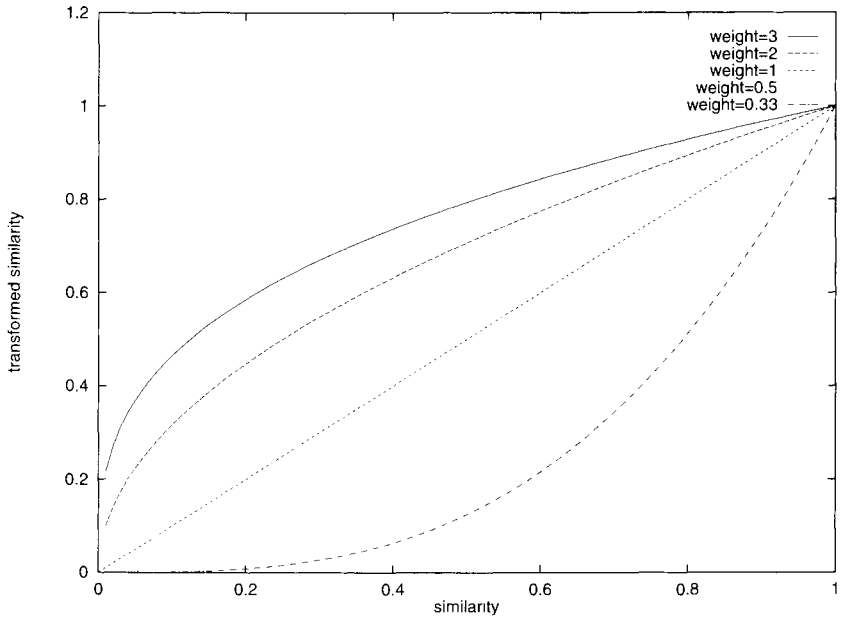


Figure 3. Various Samples for Similarity Mappings.

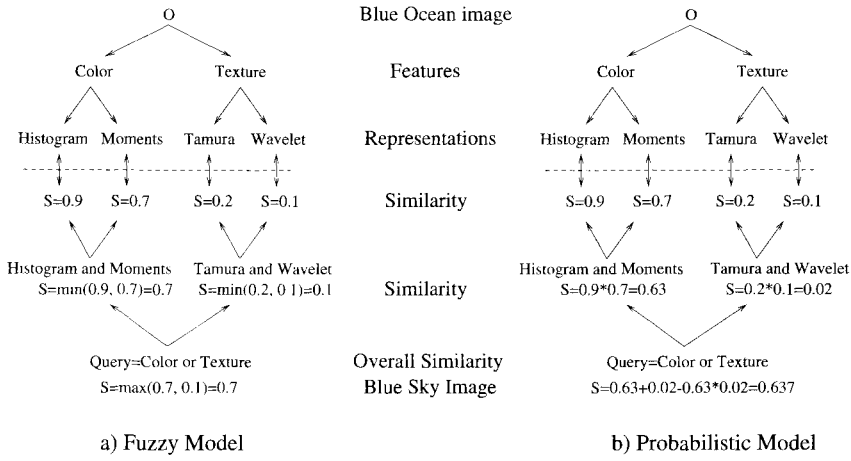


Figure 4. Various Samples for Similarity Mappings.

Another approach to optimizing query processing over multimedia repositories has been proposed in Chaudhari and Gravano (1996). It presents a strategy to optimize queries when users specify thresholds on the grade of match of acceptable objects as filter conditions. It uses the results in Fagin (1996) to convert top-k queries to threshold queries and then process them as filter conditions. It shows that, under certain conditions (uniquely graded repository), this approach is expected to access no more objects than the strategy in Fagin (1996). Furthermore, while the above approaches have mainly concentrated on the fuzzy Boolean model, we consider both the fuzzy and probabilistic models in MARS. This is significant since the experimental results illustrate that the probabilistic model outperforms the fuzzy model in terms of retrieval performance, which is discussed in a later section ("Experimental Results").

Vector Model

An information retrieval model consists of a document model, a query model, and a model for computing similarity between the documents and the queries. One of the most popular IR models is the vector model (Buckley & Salton, 1995; Salton & McGill, 1983; Shaw, 1995). Various effective retrieval techniques have been developed for this model. Among these, *term weighting* and *relevance feedback* are of fundamental importance.

Term weighting is a technique for assigning different weights for different keywords (terms) according to their relative importance to the document (Shaw, 1995; Salton & McGill, 1983). If we define w_{ik} to be the weight for term t_k, $k = 1, ..., N$, in document i (D_i), where N is the number of terms. Document i can be represented as a weight vector in the term space:

$$D_i = [w_{i1}, \ldots w_{ik}, \ldots w_{iN}] \tag{4}$$

Experiments have shown that the product of *tf* (term frequency) and *idf* (inverse document frequency) is a good estimation of the weights (Buckley & Salton, 1995; Salton & McGill, 1983; Shaw, 1995). The query Q has the same model as that of a document D–i.e., it is a weight vector in the term space:

$$Q = [w_{q1}, \ldots w_{qk}, \ldots w_{qN}] \tag{5}$$

The similarity between D and Q is defined as the Cosine distance.

$$\textit{similarity}\ (\mathrm{D}, \mathrm{Q}) = \frac{\mathrm{D} \times \mathrm{Q}}{||\mathrm{D}|| \times ||\mathrm{Q}||} \tag{6}$$

where || || denotes norm-2.

As we can see from the previous subsection ("Computing Boolean Queries"), in the vector model, the specification of w_{qk}'s in Q is very critical,

since the similarity values (*similarity (D, Q)*'s) are computed based on them. However, it is usually difficult for a user to map precisely his information need into a set of terms. To overcome this difficulty, the technique of *relevance feedback* has been proposed (Buckley & Salton, 1995; Salton & McGill, 1983; Shaw, 1995). Relevance feedback is the process of automatically adjusting an existing query using information feedback by the user about the relevance of previously retrieved documents. Term weighting and relevance feedback are powerful techniques in IR. We next generalize these concepts to VIR.

Vector Query Model and Integration of Relevance Feedback to VIR

As discussed in a previous section ("The Object Model"), an object model $O(D,F,R)$, together with a set of similarity measures $M = \{m_i\}$, provides the foundation for retrieval (D,F,R,M). The similarity measures are used to determine how similar or dissimilar two objects are. Different similarity measures may be used for different feature representations. For example, Euclidean distance is used for comparing vector-based representations, while Histogram Intersection is used for comparing color histogram representations (see the earlier section on "Visual Feature Extraction").

The query model is shown in Figure 5. The query has the same form as an object, except it has weights at every branch at all levels. W_i, W_{ij}, and

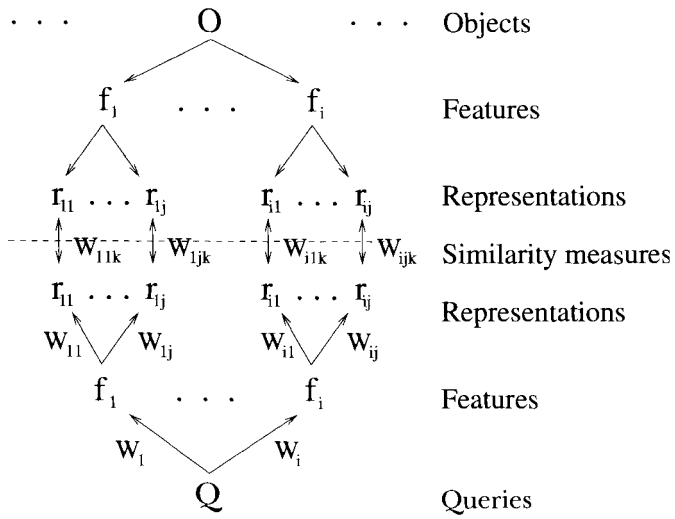


Figure 5. The Retrieval Process.

W_{ijk} are associated with features f_i, representations r_{ij}, and components r_{ijk} respectively. The purpose of the weights is to reflect as closely as possible the combination of feature representations that best express the user's information need. The process of relevance feedback described below aims at updating these weights to form the combination of features that best captures the user's information need.

Intuitively, the similarity between query and object feature representations is computed, and then the feature similarity computed as the weighted sum of the similarity of the individual feature representations. This process is repeated one level higher when the overall similarity of the object is the weighted sum over all the feature similarities. The weights at the lowest level, the component level, are used by the different similarity measures internally. Figure 6 traces this process for our familiar example of a blue sky image as a query and a blue ocean image in the collection.

Based on the image object model and the set of similarity measures, the retrieval process can be described as follows. At the initial query stage, equal weights are associated with the features, representations, and components. Best matches are then displayed back to the user. Depending on his true information need, the user will mark how good the returned matches are (degree of relevance). Based on the user's feedback, the retrieval system will automatically update weights to match the user's true information need. This process is illustrated in Figure 5. In Figure 5, the information need embedded in Q flows up while the content of O's flows

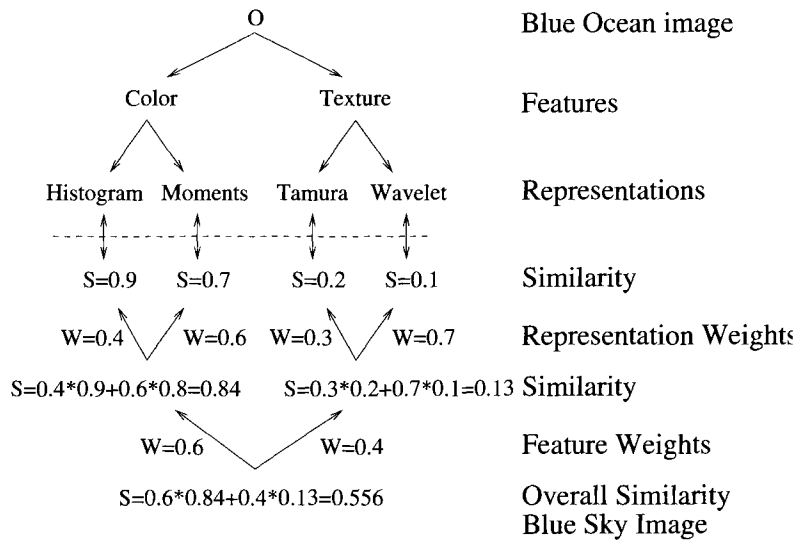


Figure 6. Example Query Calculation of Blue Sky Image against Blue Ocean Image.

down. They meet at the dashed line where the similarity measures m_{ij} are applied to calculate the similarity values $S(r_{ij})$'s between Q and O's.

Based on the intuition that important representations or components should receive more weight, we have proposed effective algorithms for updating these two levels' weights. Due to page limitation, we refer the readers to Rui et al. (1998b).

Experimental Results

In the experiments reported here, we test our approaches over the image collection from the Fowler Museum of Cultural History at the University of California—Los Angeles. It contains 286 ancient African and Peruvian artifacts and is part of the Museum Educational Site Licensing Project (MESL) sponsored by the Getty Information Institute. The size of the MESL test set is relatively small, but it allows us to explore all the color, texture, and shape features simultaneously in a meaningful way. More extensive experiments with larger collections have been performed and reported in Ortega et al. (1998b) and Rui et al. (1998b).

In the following experiments, the visual features used are color, texture, and shape of the objects in the image. The representations used are color histogram and color moments (Swain & Ballard, 1991), for the color feature Tamura (Tamura et al., 1978; Equitz & Niblack, 1994), and co-occurrence matrix (Haralick et al., 1973; Ohanian & Dubes, 1992) texture representations for the texture feature, and Fourier descriptor and chamfer shape descriptor (Rui et al., 1997b) for the shape feature.

Boolean Retrieval Model Results

To conduct the experiments, we chose several queries and manually determined the relevant set of images with the help of experts in librarianship as part of a seminar in multimedia retrieval. With the set of queries and relevant answers for each of them, we constructed precision-recall curves (Salton & McGill, 1983). These are based on the well-known precision and recall metrics. Precision measures the percentage of relevant answers, and recall measures the percentage of relevant objects returned to the user. The precision/recall graphs are constructed by measuring the precision for various levels of recall.

We conducted experiments to verify the role of feature weighting in retrieval. Figure 7(a) shows results of a *shape or color* query—i.e., to retrieve all images having either the same shape or the same color as the query image. We obtained four different precision/recall curves by varying the feature weights. The retrieval performance improves when the shape feature receives more emphasis.

We also conducted experiments to observe the impact of the retrieval model used to evaluate the queries. We observed that the fuzzy and probabilistic interpretations of the same query yield different results. Figure

7(b) shows the performance of the same query (a *texture or color* query) in the two models. The result shows that neither model is consistently better than the other in terms of retrieval.

Figure 7(c) shows a complex query (shape (I_i) *and* color (I_i) *or* shape (I_j) *and* layout (I_j)) with different weightings. The three weightings fared quite similarly, which suggests that complex weightings may not have a significant effect on retrieval performance. We used the same complex query to compare the performance of the retrieval models. The result is shown in Figure 7(d). In general, the probabilistic model outperforms the fuzzy model.

Vector Retrieval Model with Relevance Feedback Results

There are two sets of experiments reported here. The first set of experiments is on the efficiency of the retrieval algorithm—i.e., how fast the retrieval results converge to the true results. The second set of experiments is on the effectiveness of the retrieval algorithm—i.e., how good the retrieval results are subjectively.

Efficiency of the Algorithm

As we have discussed in the section "The Object Model," the image object is modeled by the combinations of representations with their corresponding weights. If we fix the representations, then a query can be

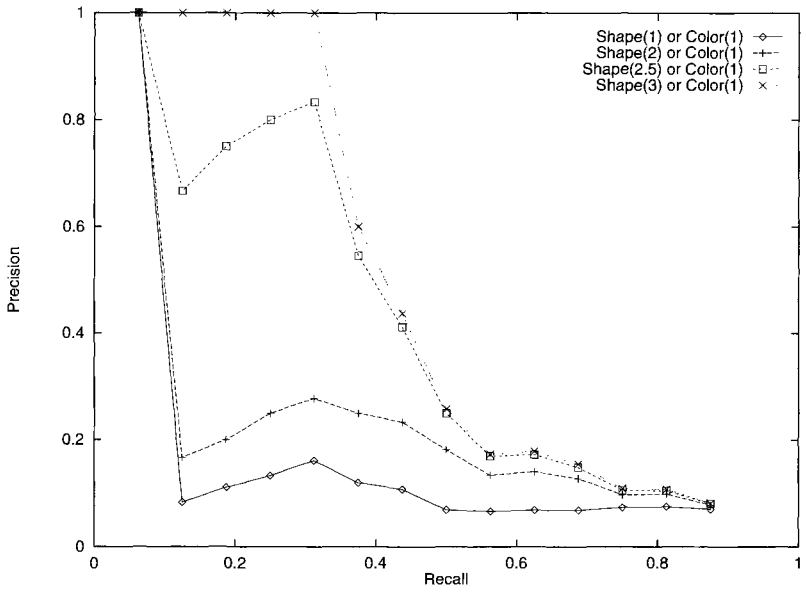
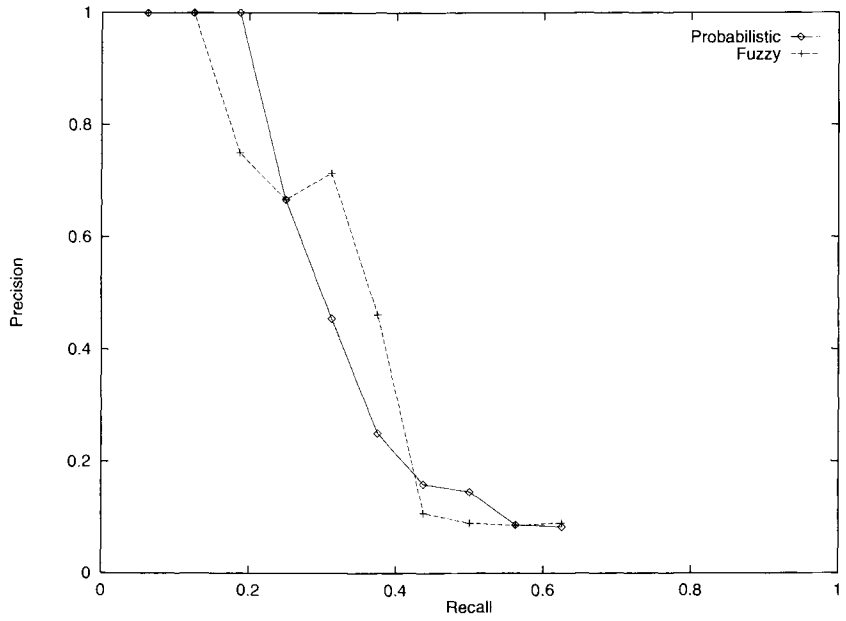
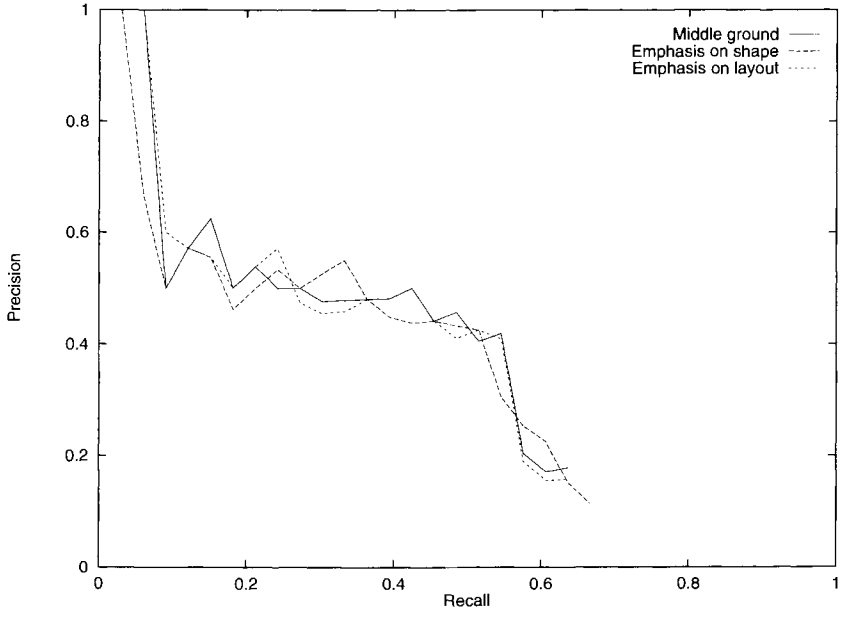


Figure 7a. Effects of Varying the Weighting on a Query.



7b. Fuzzy Versus Probabilistic Performance for Query.



7c. Complex Query with Different Weights.

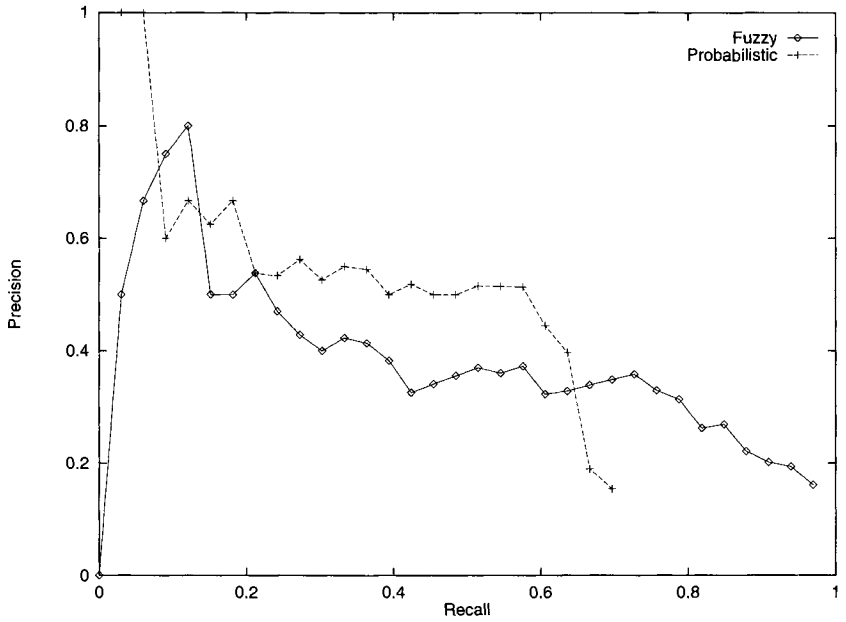


Figure 7d. Fuzzy Versus Probabilistic for SameComplex Query.

completely characterized by the set of weights embedded in the query object Q. Obviously, the retrieval performance is affected by the offset of the true weights from the initial weights. We thus classify the test into two categories—i.e., moderate offset and significant offset—by considering how far away the true weights are from the initial weights. The convergence ratio (recall) for these cases is summarized in Figure 8. Based on the curves, some observations can be made:

- In all the cases, the convergence ratio (CR) increases the most in the first iteration. Later iterations only result in minor increases in CR. This is a very desirable property, which ensures that the user gets reasonable results after only one iteration of feedback.
- CR is affected by the degree of offset. The lower the offset, the higher the final absolute CR. However, the more the offset, the higher the relative increase of CR.

Effectiveness of the Algorithm

Extensive experiments have been carried out. Users from various disciplines, such as computer vision, art, library science, and so on, as well as users from industry, have been invited to judge the retrieval performance of the proposed *interactive* approach. A typical retrieval process on the MESL test set is given in Figures 9 and 10.

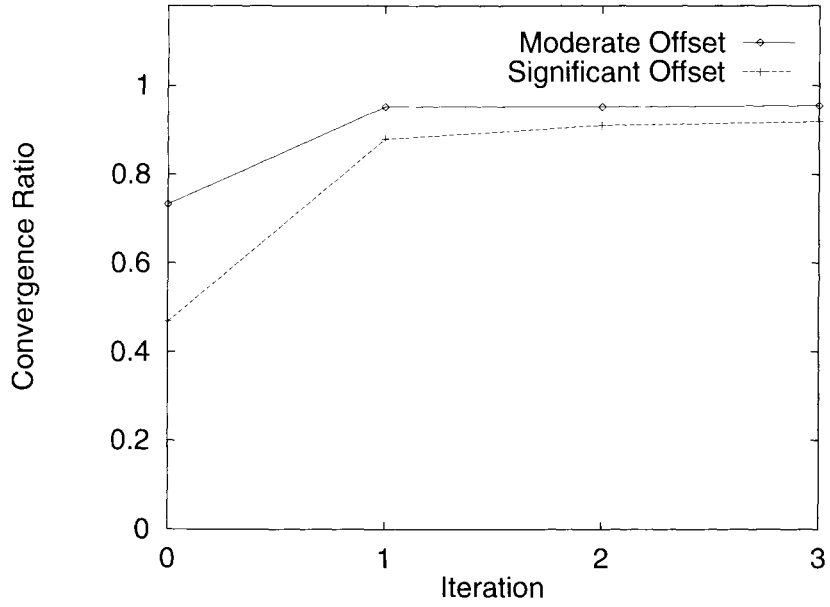


Figure 8. Convergence Ratio Curves.

The user can browse through the image database. Once the user finds an image of interest, that image is submitted as a query. In Figure 9, the query image is displayed at the upper-left corner as well as the best eleven retrieved images. The top eleven best matches are displayed in order from top to bottom and from left to right. The retrieved results are obtained based on their overall similarities to the query image, which are computed from all the features and all the representations. Some retrieved images are similar to the query image in terms of the shape feature while others are similar to the query image in terms of the color or texture feature.

Assume the user's true information need is to "retrieve similar images based on their shapes." In the proposed retrieval approach, the user is no longer required to explicitly map his or her information need to low-level features, but rather the user can express the intended information need by marking the relevance scores of the returned images. In this example, images 247, 218, 228, and 164 are marked *highly relevant.* Images 191, 168, 165, and 78 are marked *highly non-relevant.* Images 154, 152, and 273 are marked *no-opinion.*

Based on the information fed back by the user, the system *dynamically* adjusts the weights, putting more emphasis on the *shape feature,* possibly even more emphasis to one of the two shape representations which better matches the user's subjective perception of shape. The improved retrieval

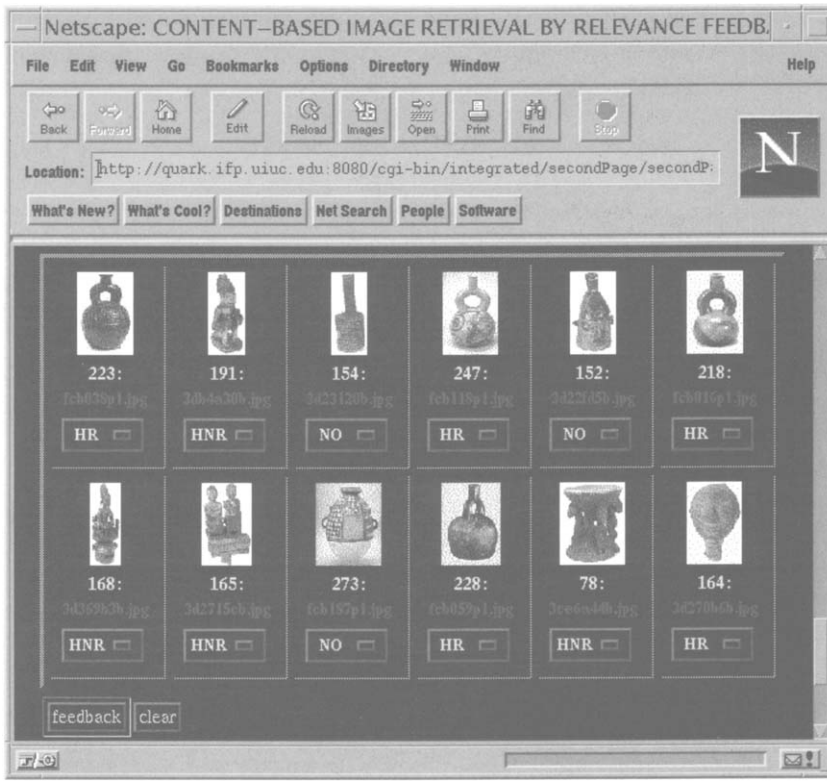


Figure 9. The Retrieval Results Before the Relevance Feedback.

results are displayed in Figure 10. Note that our shape representations are invariant to translation, rotation, and scaling. Therefore, images 164 and 96 are relevant to the query image.

Conclusion

This article discussed techniques to extend information retrieval beyond the textual domain. Specifically, it discussed how to extract visual features from images and video; how to adapt a Boolean retrieval model (enhanced with fuzzy and probabilistic concepts) for VIR systems; and how to generalize the relevance feedback technique to VIR.

In the past decade, two general approaches to VIR emerged. One is based on text (titles, keywords, and annotation) to search for visual information indirectly. This paradigm requires much human labor and suffers from vocabulary inconsistency problems across human indexers. The other paradigm seeks to build fully automated systems by completely discarding the text information and performing the search on visual information only. Neither paradigm has been very successful. In our view, these two

Figure 10. The Retrieval Results After the Relevance Feedback.

paradigms both have their advantages and disadvantages and sometimes are complimentary to each other. For example, in the MESL database, it will be much more meaningful if we first do a text-based search to confine the category and then use a visual feature-based search to refine the result. Another promising research direction is the integration of the human user into the retrieval system loop. A fundamental difference between an old pattern recognition system and today's VIR system is that the end-user of the latter is human. By integrating human knowledge into the retrieval process, we can bypass the unsolved problem of image understanding. Relevance feedback is one technique designed to deal with this problem.

Acknowledgments

This work was supported by NSF CAREER award IIS-9734300; in part by NSF CISE Research Infrastructure Grant CDA-9624396; and in part by the Army Research Laboratory under Cooperative Agreement No. DAAL01-

96-0003. Michael Ortega is supported in part by CONACYT Grant 89061 and an IBM Fellowship. Some example images used in this article are used with permission from the Fowler Museum of Cultural History at the University of California—Los Angeles.

REFERENCES

Arkin, E. M.; Chew, L.; Huttenlocher, D.; Kedem, K.; & Mitchell, J. (1991). An efficiently computable metric for comparing polygonal shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*(3), 209-216.

Bach, J. R.; Fuller, C.; Gupta, A.; Hampapur, A.; Horowitz, B.; Humphrey, R.; Jain, R.; & Shu, C-F. (1996). The Virage image search engine: An open framework for image management. In *Storage and retrieval for image and video databases IV* (Proceedings held February1-2, 1996, San Jose, CA) (pp. 76-87). Bellingham, WA: SPIE

Buckley, C., & Salton, G. (1995). Optimization of relevance feedback weights. In *SIGIR '95* (Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 9-13, 1995, Seattle, WA) (pp. 351-357). New York: Association for Computing Machinery Press.

Chaudhari, S., & Gravano, L. (1996). Optimizing queries over multimedia repositories. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data* (June 4-6, 1996, Montreal, Quebec, Canada) (pp. 91-102). New York: Association for Computing Machinery Press.

Chuang, G. C-H., & Kuo, C-C. J. (1996). Wavelet descriptor of planar curves: Theory and applications. *IEEE Transactions of Image Processing, 5*(1), 56-70.

Equitz, W., & Niblack, W. (1994). *Retrieving images from a database using texture-algorithms from the QBIC system* (IBM Computer Science Tech. Rep. No. RJ 9805). San Jose, CA: IBM.

Fagin, R. (1996). Combining fuzzy information from multiple systems. In *Proceedings of the Fifteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems* (PODS 1996, conference held June 3-5, 1996, Montreal, Canada) (pp. 216-226). New York: Association for Computing Machinery Press.

Fagin, R., & Wimmers, E. L. (1997). Incorporating user preferences in multimedia queries. In F. N. Afrati (Ed.), *Database Theory–ICDT '97* (Proceedings of the 6th International Conference, January 8-10, 1997, Delphi, Greece) (pp. 247-261). Berlin, Germany: Springer.

Flickner, M.; Sawhney, H.; Niblack, W.; Ashley, J.; Huang, Q.; Dom, B.; Gorkani, M.; Hafine, J.; Lee, D.; Petkovic, D.; Steele, D.; & Yanker, P. (1995). Query by image and video content: The QBIC system. *Computer, 28*(9), 23-32.

Haralick, R. M.; Shanmugam, K.; & Dinstein, I. (1973). Texture features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics, 3*(6), 610-621.

Kundu, A., & Chen, J-L. (1992). Texture classification using QMF bank-based subband decomposition. *Graphical Models and Image Processing, 54*(5), 369–384.

McCamy, C. S.; Marcus, H.; & Davidson, J. G. (1976). A color-rendition chart. *Journal of Applied Photographic Engineering, 2*(3), 95-99.

Miyahara, M. (1988). Mathematical transform of (R,G,B) color data to munsell (H,S,V) color data. In R. Hsing (Ed.), *Proceedings of SPIE: The Visual Society for Optical Engineering, Vol. 1001* (Visual Communications and Image Processing '88, November 9-11, 1988, Cambridge, MA) (pp. 650-657). Bellingham, WA: SPIE.

Ortega, M..; Rui, Y.; Chakrabarti, K.; Porkaew, K.; Mehrotra, S.; & Huang, T. S. (1998). Supporting ranked Boolean similarity queries in MARS. *IEEE Transactions on Knowledge and Data Engineering, 10*(6), 905-925.

Pentland, A.; Picard, R. W.; & Sclaroff, S. (1996). Photobook: Content-based manipulation of image databases. *International Journal of Computer Vision, 18*(3), 233-254.

Rui, Y.; She, A. C.; & Huang, T. S. (1996). Modified Fourier descriptors for shape representation—a practical approach. In A. Smeulders & R. Jain (Eds.), *Image databases and multi media search* (pp. 165-180). River Edge, NJ: World Scientific.

Rui, Y.; Huang, T. S.; Ortega, M.; & Mehrotra, S. (1998). Relevance feedback: A power tool in interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology, 8*(5), 644-655.

Rui, Y.; Huang, T. S.; & Chang, S-F. (1999). Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communication and Image Representation, 10*(1), 39-62.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval.* New York: McGraw-Hill Book Company.

Shaw, W. M. (1995). Term-relevance computations and perfect retrieval performance. *Information Processing and Management, 31*(4), 491-498.

Smith, J. R., & Chang, S-F. (1996). Tools and techniques for color image retrieval. In *Storage & retrieval for image and video databases IV* (Proceedings of the International Society for Optical Engineering, vol. 2670) (pp. 426-437). Bellingham, WA: SPIE.

Stricker, M., & Orengo, M. (1995). Similarity of color images. In W. Niblack & R. C. Jain (Eds.), *Storage and retrieval for image and video databases III* (Proceedings of the International Society for Optical Engineering, vol. 2420) (pp. 381-392). Bellingham, WA: SPIE.

Swain, M., & Ballard, D. (1991). Color indexing. *International Journal of Computer Vision, 7*(1), 11-32.

Tamura, H.; Mori, S.; & Yamawaki, T. (1978). Texture features corresponding to visual perception. *IEEE Transactions on Systems, Man, and Cybernetics, 8*(6), 460-473.

Additional References

Hu, M. K. (1962). Visual pattern recognition by moment invariants, computer methods in image analysis. In *IRE Transactions on Information Theory* (316 p.). New York: Institute of Radio Engineers.

Ortega, M.; Chakrabarti, K.; Porkaew, K.; & Mehrotra, S. (1998). *Cross media validation in a multimedia retrieval system.* Unpublished paper presented at the ACM Digital Libraries '98 Workshop on Metrics in Digital Libraries.

Ortega, M.; Rui, Y.; Chakrabarti, K.; Mehrotra, S.; & Huang, T. S. (1997). Supporting similarity queries in MARS. In *Proceedings of ACM Multimedia '97* (November 9-13, 1997, Seattle, WA) (pp. 403-413). New York: Association for Computing Machinery.

Rui, Y.; Huang, T. S.; & Mehrotra, S. (1997). Content-based image retrieval with relevance feedback in MARS. In *Proceedings of the International Conference on Image Processing* (October 26-29, 1997, Santa Barbara, CA) (pp. 815-818). Los Alamitos, CA: IEEE Computer Society.

Rui, Y.; Huang, T. S.; & Mehrotra, S. (1998). Exploring video structure beyond the shots. In *Proceedings of the International Conference on Multimedia Computing and Systems* (June 28-July 1, 1998, Austin, TX) (pp. 237-240). Los Alamitos, CA: IEEE Computer Society.

Precise and Efficient Retrieval of Captioned Images: The MARIE Project

NEIL C. ROWE

ABSTRACT

THE MARIE PROJECT HAS EXPLORED knowledge-based information retrieval of captioned images of the kind found in picture libraries and on the Internet. It exploits the idea that images are easier to understand with context, especially descriptive text near them, but it also does image analysis. The MARIE approach has five parts: (1) find the images and captions; (2) parse and interpret the captions; (3) segment the images into regions of homogeneous characteristics and classify them; (4) correlate caption interpretation with image interpretation using the idea of focus; and (5) optimize query execution at run time. MARIE emphasizes domain-independent methods for portability at the expense of some performance, although some domain specification is still required. Experiments show MARIE prototypes are more accurate than simpler methods, although the task is very challenging and more work is needed. Its processing is illustrated in detail on part of an Internet World Wide Web page.

INTRODUCTION

Multimedia data are increasingly important information resources for computers and networks. Much of the excitement over the World Wide Web is about its multimedia capabilities. Images of various kinds are its most common nontextual data. But finding the images relevant to some user need is often much harder than finding text for a need. Careful content analysis of unrestricted images is slow and prone to errors. It helps to find captions or descriptions as many images have them.

Neil C. Rowe, Department of Computer Science, Code CS/Rp, U. S. Naval Postgraduate School, Monterey, CA 93943
LIBRARY TRENDS, Vol. 48, No. 2, Fall 1999, pp. 475-495

Nonetheless, multimedia information retrieval is still difficult. Many problems must be solved just to find caption information in the hope of finding related images. Only a 1 percent success rate was obtained in experiments trying to retrieve photographs depicting single keywords like "moon" and "hill" using the AltaVista search engine on the World Wide Web (Rowe & Frew, 1998). This is probably because most text on Web pages, and even on pages with well-captioned photographs, was irrelevant to the photographs, and the words searched for had many senses. To improve this performance several things are needed:

- a theory of where captions are likely to be on pages;
- a theory of which images are likely to have described content;
- language-understanding software to ascertain the correct word senses and their interrelationships in the caption candidates;
- image-understanding software to obtain features not likely to be in the caption;
- a theory connecting caption concepts to image regions; and
- efficient methods of retrieval of relevant images in response to user queries or requests.

Note that speed is only critical in the last phase; while efficient methods are always desirable, accuracy is more a concern because it is so low for keyword-based retrieval. The time to do careful language and image processing can be justified during the indexing of a database if it can significantly improve later retrieval accuracy.

Consider Figure 1, which shows part of a U.S. Army Web page. Much text is scattered about, but not all of it refers to the pictures. The two formal captions (in italics, but not otherwise identified) are inconsistently placed with respect to their photographs. But the title "Gunnery at Udairi" is a caption too. Next, note that many of the words and phrases in these candidate captions do not describe the pictures. Neither picture shows "U.S. Army Central Command," "power generator equipment," an "Iraqi," "the Gulf War," or "Fort Hood"; matching would falsely retrieve this page for any of these key phrases. Similarly, the words "commander," "senior," "signal," "fire," "target," and "live" are all used in special senses, so this page would be falsely retrieved for queries intending to refer to their most common senses. The only way to eliminate such errors is to parse and interpret caption candidates using detailed linguistic knowledge. Finally, note that many things seen in both photographs are not mentioned in their captions. Only a small part of the left photograph area is devoted to the people, and the right photograph displays many features of the tanks not mentioned in its caption. Thus there are many challenges in indexing the multimedia data on these pages.

Noteworthy current systems for image retrieval are QBIC (Flickner et al., 1995), Virage (Virage Inc., San Mateo, California, USA), and

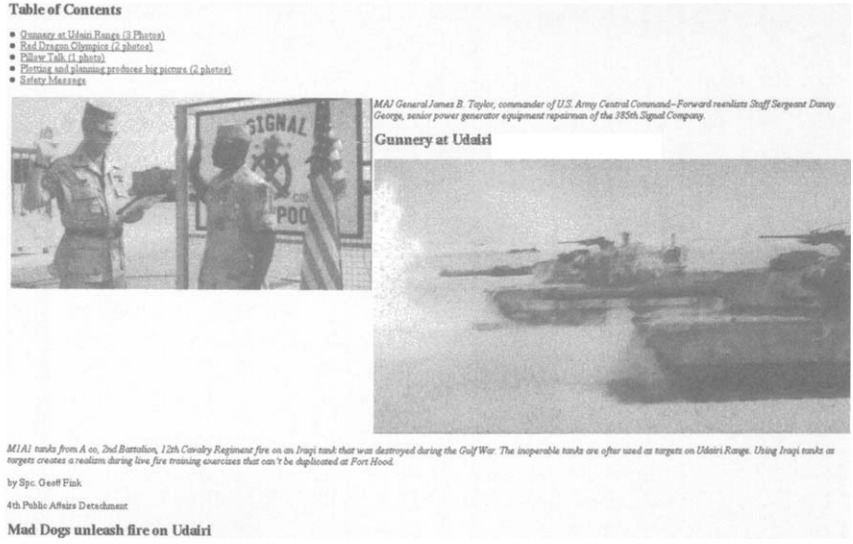


Figure 1. Example Portion of a World Wide Web Page.

VisualSEEK (Smith & Chang, 1996), which exploit simple-to-compute image properties like average color and texture. The user specifies color and texture patches, or perhaps an entire image, which is then compared to the images in the database to find the best matches. But these systems strongly emphasize visual properties and can confuse very different things of accidentally similar appearance, like seeing a face in an aerial photograph. So these systems would not help for a typical Web page like Figure 1 since color similarity to the images there would not mean much. Another category of current image-retrieval systems like Chabot (Ogle, 1995) primarily exploits descriptive information about the image, but all this information must be entered manually for each image by someone knowledgeable about it, which requires a considerable amount of tedious work.

The most interesting current research has focused on knowledge-based multimodal methods for addressing the limitations of current systems. Work on indexing of video (Hauptman & Witbrock, 1997; Smoliar & Zhang, 1994) has achieved success using knowledge-based multimodal analysis of images, image-sequence segmentation, speech, on-screen text, and closed-caption information. For single-image recognition, Piction (Srihari, 1995) does natural-language understanding of the caption of an image, and combines this information with results of face localization in the image, to provide a deeper understanding of an image. But Piction assumed that captions were already isolated for each image, and there are many interesting image features besides faces.

This article summarizes a promising approach that the MARIE project has explored recently using knowledge-based methods for accurate photograph retrieval in response to English queries by a user. The idea is to consider image retrieval in a broader perspective than that of Piction. The subtasks are finding the image, analyzing all relevant text, analyzing the image, mapping the results of the text analysis to the results of the image analysis, and efficient subsequent retrieval of this information. By considering these subtasks as parts of a larger context, we will see important issues not addressed by piecemeal efforts.

The methods of MARIE were tested in three partial prototype systems of MARIE-1, MARIE-2, and MARIE-3. These systems primarily address photographs since many users consider them the most valuable multimedia objects, but most of the methods generalize easily. Both explicit photograph libraries (especially the Photo Lab of NAWC-WD, China Lake, California, USA, with its images depicting a wide range of activities at a naval aircraft test facility) and implicit libraries (especially the World Wide Web) were investigated. Most of the code is in Quintus Prolog with some key sections in C. The remainder of the article discusses in turn each of the main problems that MARIE faced.

Locating Indexable Images and Their Captions

Identifying images and their captions is a significant problem with book-like multimedia data (as Figure 1). Web images are easy to identify by the HTML page-markup language used. But symbolic graphics, of no value to index, are generally stored the same way as photographs, as files in GIF or JPEG format, so a useful system must distinguish them. Recent work with the MARIE-3 system (Rowe & Frew, 1998) has shown that seven quickly-found parameters of images are sufficient to distinguish photographs with 70 percent recall (a fraction of all photographs found) and 70 percent precision (a fraction of items found that are photographs) on a test set of random Web pages. The parameters are size, squareness, number of distinct colors, fraction of pure colors (white, black, pure gray, red, green, and blue), color variation between neighbor pixels, variety of colors, and use of common nonphotograph words (like "button" or "logo") in the name on the image file. The parameters are converted to probabilities by applying "sigmoid" (S-shaped) functions of the form yi = tanh ((xi - mu) / sigma) where mu and sigma are constants chosen to set the center and steepness of the sigmoid curve. The probabilities are then input to a "linear-classifier neuron" calculating w0 + w1 y1 + w2 y2 + ... + w7 y7 for a set of weight constants wi determined by training. If the calculation results in a positive number, the image is considered a photograph. For Figure 1, MARIE-3 rated the left image as 0.244 and the right image as 0.123 after training, so both were correctly classified as photographs.

MARIE-3 then looks for captions around each possible photograph.

Captions are not often easy to identify because they take many forms. It is best to work on the HTML source code of the Web page, parsing it to group related things. Another seven-input linear-classifier neuron with sigmoid functions on its inputs can rate the caption candidates. Its input parameters were easy-to-calculate properties of text: distance in lines from the candidate caption to the image reference, number of other candidates at the same distance, strength of emphasis (e.g., italics), appropriateness of candidate length, use of common (counted positively) or uncommon (counted negatively) words of captions, number of identical words between candidate and either image file name or its nongraphics substitute, and fraction of the words having at least one physical-object sense. Figure 2 shows the caption candidates for Figure 1 with their ratings.

The caption neuron by itself showed 21 percent recall for 21 percent precision in matching caption-image pairs. This can be improved by combining its rating with the photograph rating since photographs are much more likely to have captions. Neuron outputs can be converted to probabilities and multiplied, with three refinements. Since a survey of random Web photographs showed that 7 percent had no visible captions, 57 percent had one caption, 26 percent had two, 6 percent had three, 2 percent had four, and 2 percent had five, it is reasonable to limit each image to its three best captions. If a caption can go with more than one image, rule out everything but the strongest match since a useful caption should be precise enough to describe only one image. For example, the "MAJ General" caption candidate in Figure 1 goes better with the left image "image4" (since captions are more often below than above) so the match to "image1" was ruled out; and "Gunnery at Udairi" goes better with the right picture from examination of the HTML code even though it is displayed similarly to the "MAJ General" candidate. Finally, consider possible "invisible" captions—the image file name, the name of any Web page pointed to by the image, any text-equivalent string for the image, and the Web page title—when their likelihoods exceed a threshold.

All this gave 41 percent recall with 41 percent precision on a random test set, or 70 percent recall with 30 percent precision, demonstrating the value of multimodal evidence fusion. Processing required 0.015 cpu seconds per byte of HTML source code, mostly in the image analysis, and the program consisted of 83 kilobytes of source code. Figure 3 shows the final captions found by MARIE-3 for Figure 1, one for the first photograph and two for the second.

Linguistic Processing

Lexical Processing

Once likely captions are found, their words can be indexed as keywords for later retrieval (excluding words that are not nouns, verbs,

Referred Image	*Caption Type*	*Caption Distance*	*Caption Candidate*
image4	plaintext	-1	"gunnery at udairi range (3 photos)"
image4	plaintext	-1	"red dragon olympics (2 photos)"
image4	plaintext	-1	"pillow talk (1 photo)"
image4	plaintext	-1	"plotting and planning produces big picture (2 photos)"
image4	plaintext	-1	"safety message"
image4	bold	-2	"table of contents"
image4	emphasis	1	"maj general james b. taylor, commander of u.s. army central command-forward reenlists staff sergeant danny george, senior power generator equipment repairman of the 385th signal company."
image4	heading2	2	"gunnery at udairi"
image1	heading2	0	"gunnery at udairi"
image1	emphasis	-1	"maj general james b. taylor, commander of u.s. army central command-forward reenlists staff sergeant danny george, senior power generator equipment repairman of the 385th signal company."
image1	emphasis	1	"m1a1 tanks from a co, 2nd battalion, 12th cavalry regiment fire on an iraqi tank that was destroyed during the gulf war."
image1	emphasis	1	"the inoperable tanks are often used as targets on udairi range."
image1	emphasis	1	"using iraqi tanks as targets creates a realism during live fire training exercises that can't be duplicated at fort hood."
image1	plaintext	3	"4th public affairs detachment"

Figure 2. Candidate Captions Inferred by MARIE-3 for the Web Page in Figure 1.

Image	*Caption*	*Final rating*
image4	"MAJ general james b. taylor, commander of u.s. army central command-forward reenlists staff sergeant danny george, senior power generator equipment repairman of the 385th signal company."	0.937
image1	"Gunnery at Udairi"	0.954
image1	"m1a1 tanks from a co, 2nd battalion, 12th cavalry regiment fire on an iraqi tank that was destroyed during the gulf war."	0.929

Figure 3. Final captions inferred by MARIE-3 for the Web page in Figure 1, with their final ratings.

adjectives, or adverbs). Several information-retrieval systems and Web search tools do this. But retrieval precision, as a result, will not be high because a word can have many meanings and many relationships to neighboring words. Nonetheless, most captions are unambiguous in their context. "View of planes on taxi from tower" is unambiguous in our NAWC-WD Navy-base test captions since "planes" are always aircraft and not levels, "tower" is always a control tower when aircraft are mentioned, "taxi" has a special meaning for aircraft, aircraft are always taxiing and not on top of a taxicab, and the view (not the aircraft or taxiing) is from the tower. Keyword-based retrieval will thus get many incorrect retrievals with the words of this caption. They would furthermore usually miss captions having synonyms or generalizations of the words, like for the caption "Photograph of 747's preparing to takeoff as seen from control" and the query "View of planes on taxi from tower."

Fortunately, true caption language understanding is easier than most text understanding (like automatic indexing of journal articles) since captions must describe something visible. Caption language heavily emphasizes physical objects and physical actions; verbs usually appear as participles, with a few gerunds and past tenses; and words for social interactions, mental states, and quantifications are rare. All this simplifies analysis. Also, many valuable applications involve technical captions, whose accessibility could be valuable to enhance, but whose difficulty primarily resides in code words and unusual word senses that are nonetheless unambiguous, grammatically easy to classify, and often defined explicitly somewhere (e.g., "zeppo" of "zeppo radar").

Many caption words are familiar words of English, and MARIE's parser needs only their parts of speech and superconcepts, obtained from the Wordnet thesaurus system (Miller et al., 1990). For the remaining words, caption writers often try to be clear and follow simple recognizable lexical

rules like "F-" followed by a number is a fighter aircraft and any number followed by "kg" is a weight in kilograms. In developing MARIE-2, such rules covered 17,847 of the 29,082 distinct words occurring in 36,191 NAWC-WD captions (see Figure 4). Person names, place names, and manufacturer names were obtained in part from existing databases for a total of 3,622 words. Of the remaining words, 1,174 were misspellings, of which 773 were correctly deciphered by our misspelling-detection software (including misspellings of unknown words, by examining word frequencies). Of the remaining words, 1,093 were abbreviations or acronyms, of which 898 were correctly deciphered by our abbreviation-hypothesizing and misspelling-fixing software (Rowe & Laitinen, 1995) using context and analogy. Of the remaining equipment names, 1,876 were not important to define further. That left 1,763 words needing explicit definition; almost

Number of captions	36,191
Number of words in the captions	610,182
Number of distinct words in the captions	29,082
Subset having explicit entries in Wordnet	6,729
Number of word senses given for these words	14,676
Subset with definitions reusable from MARIE-1	770
Subset that are morphological variants of other known words	2,335
Subset that are numbers	3,412
Subset that are person names	2,791
Subset that are place names	387
Subset that are manufacturer names	264
Subset that have unambiguous defined-code prefixes	3,256
Unambiguous defined-code prefixes	947
Subset that are other identifiable special formats	10,179
Subset that are identifiable misspellings	1,174
Misspellings found automatically	713
Subset that are identifiable abbreviations	1,093
Abbreviations found automatically	898
Subset with definitions written explicitly for MARIE-2	1,763
Remaining words, assumed to be equipment names	1,876
Explicitly used Wordnet alias facts of above Wordnet words	20,299
Extra alias senses added to lexicon beyond caption vocabulary	9,324
Explicitly created alias facts of above non-Wordnet words	489
Other Wordnet alias facts used in simplifying the lexicon	35,976
Extra word senses added to lexicon beyond caption vocabulary	7,899
Total word senses handled (includes related superconcepts, wholes, and phrases)	69,447

Figure 4. Statistics on the MARIE-2 lexicon for the NAWC-WD captions.

all of these were nouns. MARIE-2's 29,082-word lexicon construction required only 0.4 of a man-year, and much of the work can be reused unchanged for other technical applications. So porting MARIE-2 requires some work but not much.

Statistical Parsing

Although captions are usually unambiguous in their subject areas, effort is needed to determine the word senses and word relationships used in a subject area because many of these are atypical of natural language. This information could be laboriously defined manually for each subject area, but a better way is to learn it from context. This can be done with a statistical parser that learns word-sense frequencies and sense-association frequencies from a corpus of examples. A bottom-up chart parser (Charniak, 1993) will suffice. This is natural-language processing software that builds up interpretations of larger and larger substrings of the input word list by always trying to combine the strongest substring interpretations found up to that point. The strength of an interpretation can reflect the relative frequencies of the word senses used, the frequencies of the word-sense associations made, and the frequencies of the parse rules used.

To simplify matters, restrict the grammar to unary and binary parse rules (rules with only one or two symbols as the replacement for some grammatical symbol). The degree of association between two word strings in a binary parse rule can be taken as the degree of association of the two headwords of the strings. Headwords are the subjects of noun phrases, the verbs of verb phrases, sentences, and clauses, the prepositions of prepositional phrases and so on. So the headword of "F-18 aircraft on a runway" would be "aircraft," and the headword of "on a runway" would be "on." Then a particular caption interpretation can be rated by combining the degrees of association of the headword pairs at every node of its parse tree with a priori frequencies of the word senses for leaves of the tree. This is a classic problem in evidence fusion, and the simplest solution is to assume independence and multiply the probabilities. It also helps to weight the result by an adjustable monotonically-increasing function of the sentence length to encourage work on longer subphrases.

However, word-sense association frequencies are often sparse and statistically unreliable. So use "statistical inheritance" to estimate frequencies from more general ones. For instance, to estimate the frequency of "on" as a preposition and "runway" as the subject of its prepositional phrase, look for statistics of "on" and any surface or, if the sample size is not enough, "on" and a physical object. When sufficiently reliable frequency statistics for a modified pair are found, divide by the ratio of the number of occurrences of the substitute word to the number of occurrences of "runway," since a more general concept should associate proportionately more with

"on." Similarly, generalize "on" to the class of all physical-relationship prepositions, or generalize both "on" and "runway" simultaneously. Generalize even to the class of all prepositions and the class of all nouns (i.e., the parse-rule frequency) if necessary, but those statistics are less reliable; one should prefer the minimum generalization having statistically reliable data. Statistical inheritance is self-improving because each confirmed interpretation provides more data.

This statistical approach to parsing addresses the two main linguistic challenges of captions—interpretation of nouns modifying other nouns ("nominal compounds") and interpretation of prepositional phrases. Both involve inference of case relationships. Figure 5 lists important cases for nominal compounds in the captions studied in Rowe and Frew (1998). Distinguishing these cases requires a good type hierarchy which Wordnet can supply as well as providing synonym and part-whole information for word senses.

Example	*Case relationship*
B-747 aircraft	subtype-type
aircraft wing	whole-part
F-18 Harrier	object-alias
aircraft BU#194638	type-identifier
lights type iv	object-type
rock collection	object-aggregation
mercury vapor	material-form
10-ft pole	measure-object
infrared sensor	mode-object
wine glass	associate-object
Delta B-747	owner-object
Captain Jones	title-person
Commander NWC	job-organization
aircraft closeup	object-view
foreground clouds	view-object
Monterey pharmacy	location-object
sea site	object-location
NRL Monterey	organization-location
Benson Arizona	sublocation-location
assembly area	action-location
arena test	location-action
reflectivity lab	subject-location
assembly start	action-time
July visit	time-action
training wheels	action-object
parachute deployment	object-action
air quality	object-property
project evaluation	concept-concept
night-vision goggles	concept-object
ECR logo	object-symbol

Figure 5. Important Cases of Nominal Compounds for Captions.

The result of parsing and semantic interpretation of a caption is a meaning representation. Since captions so rarely involve quantification, tenses, and hypothetical reasoning, their meaning is generally expressible with "conjunctive semantics"—as a list of type and relationship facts. Each noun and verb maps to an instance of its word-sense type, and instances are related with relationship predicates. Figure 6 shows the semantic interpretations found by MARIE-3 for the inferred captions of Figure 1. Multiple captions found for the same image were treated as separate sentences. The "v" symbols are existentially quantified variables for type instances, and hyphenated numbers are sense numbers. Sense

Inferred caption on left photograph: "maj general james b. taylor, commander of u.s. army central command-forward, reenlists staff sergeant danny george, senior power generator equipment repairman of the 385th signal company."

[a_kind_of(v435,enlist-103), quantification(v435,singular),
property(v435,repeated), object(v435,v14),
a_kind_of(v14, Staff Sergeant-0), a_kind_of(v14, George-0), identification(v14, Danny),
a_kind_of(v14,serviceman-2), subject(v14,v558),
a_kind_of(v558,equipment-1), part_of(v558,v486),
a_kind_of(v486,generator-2), agent(v464,v486), a_kind_of(v464,power-3),
property(v14,senior-51), owned_by(v14,v22),
a_kind_of(v22,Signal Company-0), property(v22,385th-50),
rank(v22,385), quantification(v22,the), agent(v435,v11),
a_kind_of(v11,Major General-0), a_kind_of(v11,James-0),
a_kind_of(v11,Taylor-0), identification(v11,B.-0),
a_kind_of(v11,commander-2), located_at(v11,v212),
a_kind_of(v212,front-1), located_at(v212,v209), a_kind_of(v209,command-0),
property(v212,central-52), part_of(v212,v8),
a_kind_of(v8,United States Army -0)]).

Inferred caption on right photograph: "gunnery at udairi. m1a1 tanks from a co, 2nd battalion, 12th cavalry regiment, fire on an iraqi tank that was destroyed during the gulf war."

[a_kind_of(v1,gunnery-1), at(v1,v15), a_kind_of(v15, Udairi -0), during(v1,v18),
a_kind_of(v18,shoot-109), quantification(v18,plural), object(v18,v148),
a_kind_of(v148,tank-4), property(v148, Iraqi -50),
quantification(v148,a), object(v156,v148),
a_kind_of(v156,destroy-101), tense(v156,past),
quantification(v156,singular), during(v156,v21),
a_kind_of(v21, Gulf War -0), quantification(v21,the), agent(v18,v23),
a_kind_of(v23, M1A1 -0), quantification(v23,plural), from(v23,v49),
a_kind_of(v49,company-4), quantification(v49,a), owned_by(v49,v75),
a_kind_of(v75,battalion-1), property(v75, 2nd -51), owned_by(v75,v89),
a_kind_of(v89,regiment-1), subject(v89,v85),
a_kind_of(v85,cavalry-1), property(v89, 12th -50), rank(v89,12)]).

Figure 6. Semantic Interpretations Found by MARIE-3 for the Inferred Captions of Figure 1.

numbers 0-49 are for nouns (with nonzero sense numbers from Wordnet version 1.5); 50-99 are for adjectives (with the number minus 50 being the Wordnet sense number); and 100-149 are for verbs (with the number minus 100 being the Wordnet sense number). The "a_kind_of" expressions relate a variable to a type, so for instance "a_kind_of(v435,enlist-103)" says the caption refers to some v435 that is an instance of the verb "enlist" in Wordnet sense number 3. Other expressions give properties of variables, like "quantification(v435,singular)" meaning that the enlisting event was singular as opposed to plural, and two-variable expressions relate the variables, like "object (v435,v14)" meaning that the object of the enlisting event was some v14, an instance of "Staff Sergeant."

In general, MARIE-2 (MARIE-3 is not finished) took a median CPU time of 9.7 seconds (and a geometric mean of 10.2 seconds, the antilogarithm of the mean of the logarithms) to parse randomly selected NAWC-WD caption sentences, sentences averaging 7.4 words in length, with more difficulty for a few sentences with rare syntactic constructs. This processing used 226K of source code and 6832K of data (including 1894K of word sense statistics and 1717K of binary statistics). MARIE-2 found the correct interpretation on its first try for the majority of the test sentences, with a geometric mean of 1.9 tries. Figure 7 shows the value of different kinds of linguistic information to the interpretation of some representative sentences. Figure 8 shows statistics on four successive test sets on this particular technical dialect; word-frequency statistics for each set were calculated before going to the next, so "learning" only took place then. Note how the introduction of new syntactic and semantic rules is declining, although significant numbers of new words are still being introduced in this open-ended real-world dialect.

The caption's meaning representation can be indexed in a database under the picture name. The same linguistic processing methods can interpret natural-language queries, look up word senses in the index, and do a subgraph-graph isomorphism match between the query semantic network and the caption semantic network (Rowe, 1996; Guglielmo & Rowe, 1996) to prove that the query is covered by the caption.

Image Processing

While captions are generally simpler to analyze than unrestricted natural language, captioned images are not much easier than most images; they are the most valuable when they show much variety. Linguistic processing also takes much less time than image processing; NAWC-WD captions average twenty-two words long while their pictures need 10,000 pixels for minimal representation. Nonetheless, image processing of captioned images can provide valuable information not in captions. Captions rarely describe the size, orientation, or contrast of important objects in the image. They rarely describe easy-to-see features like whether the

Number	*Sentence*
1	pacific ranges and facilities department, sled tracks.
2	airms, pointer and stabilization subsystem characteristics.
3	vacuum chamber in operation in laser damage facility.
4	early fleet training aid: sidewinder 1 guidance section cutaway.
5	awaiting restoration: explorer satellite model at artifact storage facility.
6	fae i (cbu-72), one of china lake's family of fuel-air explosive weapons.
7	wide-band radar signature testing of a submarine communications mast in the bistatic anechoic chamber.
8	the illuminating antenna is located low on the vertical tower structure and the receiving antenna is located near the top.

Sentence number	*Training*		*Final*		*No binary*		*No unary*	
	Time	*Tries*	*Time*	*Tries*	*Time*	*Tries*	*Time*	*Tries*
1	27.07	13	17.93	5	8.27	5	60.63	19
2	70.27	10	48.77	9	94.62	14	124.9	23
3	163.0	19	113.1	19	202.9	23	2569.0	22
4	155.2	9	96.07	3	63.95	8	229.3	22
5	86.42	8	41.02	3	49.48	6	130.6	30
6	299.3	11	65.78	7	68.08	5	300.4	15
7	1624.0	24	116.5	5	646.0	12	979.3	25
8	7825.0	28	35.02	2	35.60	3	>50000	-

Figure 7. Example sentences and their interpretation times in CPU seconds during training; after training; after training without binary co-occurence frequencies; and after training without unary word-sense frequencies.

Statistic	*Training set 1*	*Training set 2*	*Training set 3*	*Training set 4*
Number of new captions	217	108	172	119
Number of new sentences	444	219	218	128
Number of total words in new captions	4488	1774	1535	1085
Number of distinct words in new captions	939	900	677	656
Number of new lexicon entries required	c.150	106	139	53
Number of new word senses used	929	728	480	416
Number of new sense pairs used	1860	1527	1072	795
Number of lexical-processing changes required	c.30	11	8	7
Number of syntactic-rule changes or additions	35	41	29	10
Number of case-definition changes or additions	57	30	16	3
Number of semantic-rule changes or additions	72	57	26	14

Figure 8. Overall Statistics on the Training Sets.

image is a daytime view, an outdoor view, or a historical photograph. Nor do they mention things obvious to people familiar with the picture subject, a serious problem for specialized technical images. For instance, captions rarely mention that sky or ground is shown in a picture, and NAWC-WD captions rarely mention that the photographs were taken at NAWC-WD.

MARIE's basic image processing segments the image into regions, computes region properties, and broadly classifies regions. It uses a robust "split-and-merge" method on thumbnail reductions (to about 10,000

pixels each) of the images. The image is split into small irregular regions based on color similarity, and the regions are merged in a best-first way using color and texture similarity until the number and size of the remaining regions simultaneously achieves several criteria. Typically this was when 50-100 regions remained. Then some final splitting criteria are applied, and seventeen key properties of the regions are calculated. The seventeen were developed from an extensive survey of a variety of captioned images. They represent key dimensions for distinguishing regions, including color, texture, density, horizontality and verticality, boundary shape, and boundary strength. Figure 9 shows the region segmentations found for the Figure 1 photographs, and Figure 10 lists properties computed for regions in general.

Reliable identification of objects in unrestricted photographs is very difficult because of the wide range of subjects and photographic conditions. Nonetheless, some experiments on the easier task of trying to classify the regions into one of twenty-five general categories (Rowe & Frew,

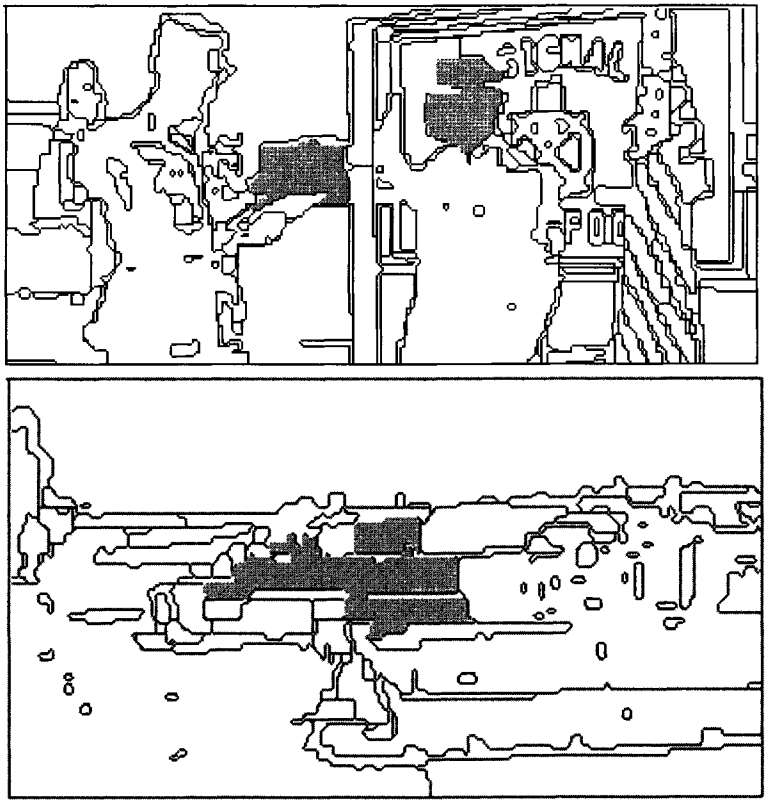


Figure 9. Segmentation and focus assignment for the photographs on the Figure 1 Web page. The shaded areas represent the computed best visual focus.

Name	Definition
circularity	area/(circumference*circumference)
narrowness	height/width of bounding rectangle
marginality	1/(1+(circumference/number border cells))
redness	average red brightness
greenness	average green brightness
blueness	average blue brightness
pixel texture	average brightness variation of adjacent cells
brightness trend	brightness correlation with x or y
symmetry	skew of center of mass from bounding-rectangle center
contrast	average strength of the region edge
diagonality	smoothed-boundary diagonality
curviness	smoothed-boundary curviness
segments	smoothed-boundary number of inflection points
rectangularity	smoothed-boundary number of right angles
size	area in pixels
density	density of pixels in bounding rectangle
height	y-skew (unsigned) of center of mass within image

Figure 10. Properties Computed for Image Regions.

1997) provide hope of helping in matching to the caption (though the next section reports a better approach). These experiments used a random sample of 128 photographs from the NAWC-WD library covering a wide variety of subjects and activities and taken under a variety of lighting conditions. A neural network was trained to classify each region of the picture as one of twenty-five—airplane, airplane part, animal, bomb, bomb part, building, building part, equipment, fire, flower, helicopter, helicopter part, missile, missile part, mountain, pavement, person, person body part, rock, ship part, sky, tank, terrain, wall, and water (some of these are domain-dependent but easily generalizable). Twenty-five classes appear sufficient for many applications because captions usually refine the classifications; if the caption mentions a B-747 and no other airplanes, that is likely to be the airplane shape in the image. The neural network takes the seventeen region properties as inputs and computes the likelihood the region is of a particular class for each of the twenty-five classes. Weights connecting inputs and outputs are learned. So the output for "equipment" exploits a high weight on the number of right angles of the region border, and the output for "sky" exploits a low weight on the brightness variation of adjacent cells.

With just this simple approach without relationship constraints, 33 percent precision was obtained in classifying the regions in a random sample. Precision was improved to 50 percent with addition of another level of neural reasoning that used the linguistic-focus ideas explained in

the next section, showing that caption information helps in image analysis. Still better performance could be achieved with appropriate domain-dependent region-relationship constraints (like that the sky is always above terrain), but domain independence is desirable for portability to other image libraries. (The classic alternative of case-based reasoning for region classification only obtained 30 percent precision for twenty-five regions; apparently some classes show too much variation in appearance.) Image processing averaged an hour per image. But again, this time is expended only during database setup when time is not critical.

Caption-Image Reference Semantics

Linguistic Focus

So far caption and image analysis have been considered separately, but their results must eventually be integrated. This requires finding the "linguistic focus" of the caption and "visual focus" of the image because these are implicitly cross-referenced. Rowe's (1994) study showed that captions are a dialect restricted in semantics as well as syntax. The headwords (usually syntactic subjects) of caption sentences usually correspond to the most important fully-visible object(s) in the image (not necessarily the largest). In conjunctions and multi-sentence captions, each headword corresponds to an important visible object. So a pod, pylon, and bracket should be clearly visible in the image for "Radar pod and pylon; front mounting bracket." Physical-action verbs are also depicted in images when appearing as gerunds, as "loading" in "Aircraft loading by crane," as participles, as past tense, or as present tense. Verbs generally correspond to relationships in the image rather than regions.

Other nouns or verbs in a caption are generally visible in part if they are related syntactically to a headword. So "Containers on truck" implies that all the containers are depicted and part but not necessarily all of the truck, while "Truck with containers" implies the opposite. Similarly, "Aircraft cockpit" guarantees only part of the aircraft is visible since "aircraft" is an adjective here. The same is true for direct objects of physical-action verbs like "resistor" in "Technician soldering resistor."

Captions also have several additional conventional forms that signal depiction, like "The picture shows X," "You can see X," and "X with Y" where "with" acts like a conjunction. Also, word forms such as "closeup of X" make X the true headword. This latter is a case of a general principle, that an undepictable headword refers its headword designation to its syntactic object. Depictability means whether a word sense is a physical object in the Wordnet concept hierarchy.

Figure 11 shows the concepts inferred by these and similar rules for the photographs of Figure 1. In tests with random photograph captions, 80 percent precision was obtained with 62 percent recall in identifying concepts shown in the photographs from the captions alone.

Another phenomenon is that some captions are "supercaptions" that refer to more than one picture, using typically-parallel syntactic constructs. For example, in "Sled test: Pretest assembly, close-up of building, parachute deployment, and post-test damage," "sled test" maps to a set of four pictures, but the four contradictory conjuncts map to each of the successive pictures. Negative depictability is also inferable when a caption does not mention something expected for its domain. For instance, NAWC-WD tests equipment for aircraft, so a test not mentioned in a caption is guaranteed not to be shown in its photograph.

Visual Focus

The linguistic focus of a caption corresponds to a subject or "visual focus" of its corresponding image. Captioned photographs alone have special visual semantics, since they are usually selected because they depict their subjects well. From a study of sample photographs, it was observed that the visual foci of captioned images were region sets with generally five characteristics:

1. they are large;
2. their center of gravity is near the picture center;
3. they do not touch the edges of the photograph;
4. their boundary has good contrast; and
5. they are maximally different from non-focus regions of the picture.

A trainable neuron was used to summarize the five factors. The best candidate region set can be found by a best-first search over region sets. This set is then matched to the linguistic focus, excluding redundant terms and those for objects too small compared to the others, like "pilot" when "aircraft" is also in linguistic focus. This requires inheritable average sizes of objects with standard deviations of their logarithms.

Performance of this approach on a random set of images (different from those tested for image processing) was 64 percent precision for 40 percent recall. These figures were computed as ratios of numbers of pixels. So, in other words, 64 percent of the pixels selected as belonging to subjects of the picture were actually part of the subjects. Precision is the challenge since 100 percent recall is easy by just designating the entire picture as the subject. Segmentation was critical for these results since only 1 percent precision was obtained by selecting all pixels whose color was significantly different from the color of any picture-boundary pixel. Altogether, caption-image reference analysis required 29K of source code and less than a second of CPU time per test caption. The shaded areas in Figure 9 represent the best hypotheses found for the subjects of the Figure 1 photographs. The program mistakenly selected a sign in the first picture that was close to the center of the image, but the other regions selected are correct as are their labels (see Figure 11). Region classifica-

image4: [enlist-103], [singular]
image4: [George -0, Staff Sergeant -0,serviceman-2], []
image4: [James -0, Major General -0, Taylor -0,commander-2], [singular]
image1: [gunnery-1], []
image1: [shoot-109], [plural]
image1: [M1A1 -0], [plural]

Figure 11. Results of linguistic focus and depiction analysis showing the terms (with any quantifications) inferred to apply to the subjects of the example photographs (the shaded areas in Figure 9).

tion in the manner of the last section helps avoid such mistakes, but it helps less than the five principles above.

Efficient Image Retrieval at Query Time

Let us now consider what happens once a set of images has been indexed by their caption and image concepts. Such stored information can be queried by parsed English queries, or also by key phrases, in the MARIE systems. Execution of such queries on a single computer processor can be thought of as sequential "information filtering" that successively rules out images on various criteria. Terms in a parse interpretation, or key phrases extracted from them, can each correspond to a separate filter, but there can be many other kinds of useful filters. Efficient query execution strategies are important in implementing these filters because speed at query time is critical to user satisfaction.

In a sequence of information filters, it often helps to put the toughest filters first to reduce workload fastest (though filter sequences are conjunctive and give the same results in any order). This is desirable when filters need a constant amount of time per data item, as in hash lookups from a sparse hash table, but not always otherwise. Rowe (1996) showed the criterion for local optimality with respect to interchange of filters i and i+1 in a filter sequence:

$$c(i) / (1 - p(f(i) \mid g(i-1))) <= c(i+1) / (1 - p (f(i+1) \mid g(i-1)))$$

Here f(i) is the event of a data item passing filter i, g(i-1) is the event of a data item passing filters 1 through i-1, c(i) is the average execution cost per data item of filter i, and p means "probability." This is only a local optimality condition since g(i-1) represents context but can be used heuristically to sort the filter sequence, and experiments showed that such sorting nearly always gave the globally optimal sequence. The criterion is especially valuable for placing information filters that do complex processing. An example is the subgraph-isomorphism check mentioned at

the end of the earlier section on "Linguistic Processing." Such matching is valuable but time-consuming, and Rowe (1996) proved it should be done last among MARIE filters.

Another way to improve the efficiency of a filter sequence is to introduce appropriate redundant information filters. Redundant filters can actually help when they are faster than the filters which make them redundant. For instance, the subgraph-isomorphism filter makes redundant the filters that only match noun senses between query and caption, but the latter can quickly cut the former's workload considerably. Another redundant filter used by MARIE-2 broadly classifies the query (for instance, into "test photo," "public relations photo," and "portrait" classes) and rules out matches with incompatible caption classes; this is also much faster but redundant with respect to the subgraph-isomorphism filter. A proven sufficient criterion for local optimality with respect to nondeletion for a redundant filter i in a filter sequence is:

$$c(i) / (1 - p (f (i) | g(i-1))) <= c(i+1)$$

with the same notation as above. Again, experiments showed this led almost always to the globally optimal sequence. Other useful analytic criteria for optimality of information filtering were shown, including those for optimality of disjunctive sequences, negations, and Boolean combinations of filters. Using such optimizations, MARIE-1 took about two seconds per query (the geometric mean of CPU time) for a sample of queries generated by actual NAWC-WD users (which were shorter and simpler than captions). MARIE-2 took about three seconds per query but gave better answers. Optimization took just a few seconds of CPU time and 23K of source code (Rowe, 1996).

Another way to speed up information filtering is by data parallelism—different processors trying different image sets to match to the query. (Other forms of parallelism do not help information filters much.) Load imbalances can occur when some processors finish early on their allocation of work because of random fluctuations in either processing time or the success rate of a filter. But data can be assigned randomly to processors, and the load imbalance estimated quite accurately at each step, which permits judging when the cost of rebalancing the processors is justified. Such parallelism would help applications requiring high-speed retrieval like real-time robots.

Conclusion

The success of text information retrieval for the Internet has obscured the considerably greater difficulty of multimedia information retrieval. Naïve methods, such as searching for keywords likely to be associated with a particular kind of image, encounter low success rates. The MARIE project has explored an integrated solution to multimedia retrieval using knowl-

edge-based methods and clues from linguistics, image analysis, presentation layout, and mathematical analysis. While perfection is not possible for information retrieval tasks, major improvements over the 1 percent success rate of naïve keyword lookup are definitely possible with the ideas presented here. Clear synergism was obtained by using multimodal clues and confirming advantages of multimodal processing (Maybury, 1997). At the same time, MARIE uses mostly domain-independent methods that are relatively easy to extend to new image libraries. The insights from MARIE may prove important in tapping the wealth of multimedia data available on the information superhighway.

Acknowledgments

This work was supported by the U. S. Army Artificial Intelligence Center and by the U. S. Naval Postgraduate School under funds provided by the Chief for Naval Operations.

References

Charniak, E. (1993). *Statistical language learning.* Cambridge, MA: MIT Press.

Flickner, M.; Sawhney, H.; Niblack, W.; Ashley, J.; Huang, Q.; Dom, B.; Gorkani, M.; Hafner, J.; Lee, D.; Petkovic, D.; Steele, D.; and Yanker, P. (1995). Query by image and video content: The QBIC System. *Computer, 28*(9), 23-32.

Guglielmo, E. J., & Rowe, N. C. (1996). Natural-language retrieval of images based on descriptive captions. *ACM Transactions on Information Systems, 14*(3), 237-267.

Hauptman, G., & Witbrock, M. (1997). Informedia: News-on-demand multimedia information acquisition and retrieval. In M. T. Maybury (Ed.), *Intelligent multimedia information retrieval* (pp. 215-239). Menlo Park, CA: AAAI Press.

Maybury, M. T. (Ed.). (1997). *Intelligent multimedia information retrieval.* Menlo Park, CA: AAAI Press.

Miller, G. A. (Ed.). (1990). WordNet: An online lexical database (special thematic issue). *International Journal of Lexicography, 3*(4), 235-312.

Ogle, V. E., & Stonebraker, M. (1995). Chabot: Retrieval from a relational database of images. *Computer, 28*(9), 40-48.

Rowe, N. C. (1994). Inferring depictions in natural-language captions for efficient access to picture data. *Information Processing and Management, 30*(3), 379-388.

Rowe, N. C. (1996). Using local optimality criteria for efficient information retrieval with redundant information filters. *ACM Transactions on Information Systems, 14*(2), 138-174.

Rowe, N. C., & Frew, B. (1997). Automatic classification of objects in captioned depictive photographs for retrieval. In M. T. Maybury (Ed.), *Intelligent multimedia information retrieval* (pp. 65-79). Menlo Park, CA: AAAI Press.

Rowe, N. C., & Frew, B. (1998). Automatic caption localization for photographs on World Wide Web pages. *Information Processing and Management, 34*(1), 95-107.

Rowe, N. C., & Laitinen, K. (1995). Semiautomatic disabbreviation of technical text. *Information Processing and Management, 31*(6), 851-857.

Smith, J., & Chang, S-F. (1996). VisualSeek: A fully automated content-based image query system. In P. Aigrain, V. Bove, W. Hall, & T. Little (Eds.), *Proceedings of ACM Multimedia '96* (November 18-22, 1996, Boston, MA) (pp. 87-98). New York: Association for Computing Machinery Press.

Smoliar, S., & Zhang, H. (1994). Content based video indexing and retrieval. *IEEE Multimedia, 1*(2), 62-72.

Srihari, R. K. (1995). Automatic indexing and content-based retrieval of captioned images. *Computer, 28*(9), 49-56.

Exploiting Multimodal Context in Image Retrieval

ROHINI K. SRIHARI AND ZHONGFEI ZHANG

ABSTRACT

THIS RESEARCH EXPLORES THE INTERACTION of textual and photographic information in multimodal documents. The World Wide Web (WWW) may be viewed as the ultimate, large-scale, dynamically changing, multimedia database. Finding useful information from the WWW without encountering numerous false positives (the current case) poses a challenge to multimedia information retrieval systems (MMIR). The fact that images do not appear in isolation, but rather with accompanying collateral text, is exploited. Taken independently, existing techniques for picture retrieval using collateral text-based methods and image-based methods have several limitations. Text-based methods, while very powerful in matching context, do not have access to image content. Image-based methods compute general similarity between images and provide limited semantics. This research focuses on improving precision and recall in an MMIR system by interactively combining text processing with image processing (IP) in both the indexing and retrieval phases. A picture search engine is demonstrated as an application.

INTRODUCTION

This research explores the interaction of textual and photographic information in multimodal documents. The World Wide Web (WWW) may be viewed as the ultimate, large-scale, dynamically changing, multi-

Rohini K. Srihari, Department of Computer Science, Center for Document Analysis and Recognition (CEDAR), UB Commons, 520 Lee Entrance—Suite 202, State University of New York, Buffalo, NY 14228-2567
Zhongfei Zhang, Computer Science Department, Watson School of Engineering and Applied Science, State University of New York at Binghamton, Vestal, NY 13902
LIBRARY TRENDS, Vol. 48, No. 2, Fall 1999, pp. 496-520

media database. Finding useful information from the WWW poses a challenge in the area of multimodal information indexing and retrieval. The word "indexing" is used here to denote the extraction and representation of semantic content. This research focuses on improving precision and recall in a multimodal information retrieval system by interactively combining text processing with image processing.

The fact that images do not appear in isolation but rather with accompanying text, which is referred to as collateral text, is exploited. Figure 1 illustrates such a case. The interaction of text and image content takes place in both the indexing and retrieval phases. An application of this research—namely, a picture search engine that permits a user to retrieve pictures of people in various contexts—is presented.

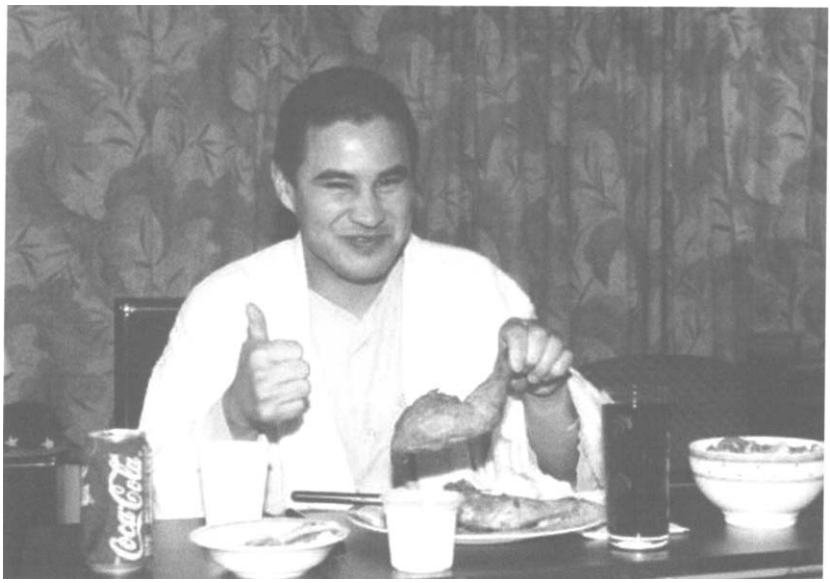


Figure 1. Staff Sgt. Andrew Ramirez digs into a plate of chicken, his first hot meal since release from captivity, at the Landstuhl Regional Medical Center, Landstuhl, Germany (U. S. Air Force photo by Senior Airman Elizabeth Weinberg). Released photo by: SRA Brian M. Boisvert, 786th Communications Squadron Record ID No. (VIRIN): 990502-F07285W-001). The picture was obtained from the U. S. Department of DefenseLink Web page located at http://defenselink.mil/multimedia/.

Taken independently, existing techniques for text and image retrieval have several limitations. Text-based methods, while very powerful in matching context (Salton, 1989), do not have access to image content. There has been a flurry of interest in using textual captions to retrieve images (Rowe & Guglielmo, 1993). Searching captions for keywords and names

will not necessarily yield the correct information, as objects mentioned in the caption are not always in the picture. This results in a large number of false positives that need to be eliminated or reduced. In a recent test, a query was posed to a search engine to find pictures of Clinton and Gore and resulted in 941 images. After applying our own filters to eliminate graphics and spurious images (e.g., white space), 547 potential pictures that satisfied the query remained. A manual inspection revealed that only 76 of the 547 pictures contained pictures of Clinton or Gore. This illustrates the tremendous need to employ image-level verification and to use text more intelligently.

Typical image-based methods compute general similarity between images based on statistical image properties (Flickner et al., 1995). Examples of such properties are texture and color (Swain & Ballard, 1991). While these methods are robust and efficient, they provide very limited semantic indexing capabilities. There are some techniques that perform object identification; however, these techniques are computationally expensive and not sufficiently robust for use in a content-based retrieval system. This is due to a need to balance processing efficiency with indexing capabilities. If object recognition is performed in isolation, this is probably true. More recently, other attempts to extract semantic properties of images based on spatial distribution of color and texture properties have also been attempted (Smith & Chang, 1996). Such techniques have drawbacks, primarily due to their weak disambiguation. These are discussed later. Webseer (http://webseer.cs.uchicago.edu) describes an attempt to utilize both image and text content in a picture search engine. However, text understanding is limited to processing of HTML tags; no attempt to extract descriptions of the picture is made. More important, it does not address the interaction of text and image processing in deriving semantic descriptions of a picture.

In this article, a system for finding pictures in context is described. A sample query would be *Find pictures of victims of natural disasters.* Specifically, experiments have been conducted to effectively combine text content with image content in the retrieval stage. Text indexing is accomplished through standard statistical text indexing techniques and is used to satisfy the general context that the user specifies. Image processing consists of face detection and recognition. This is used to present the resulting set of pictures based on various visual criteria (e.g., the prominence of faces). Experiments have been conducted on two different scenarios for this task; results from both are presented. Preliminary work in the intelligent use of collateral text in determining pictorial attributes is also presented. Such techniques can be used independently or combined with image processing techniques to provide visual verification. Thus this represents the integration of text and image processing techniques in the indexing stage.

Important Attributes for Picture Searches

Before techniques for extracting picture properties from text and images are described, it is useful to examine typical queries used in retrieving pictures. Jorgensen (1996) describes experimental work in the relative importance of picture attributes to users. Twelve high-level attributes—literal object, people, human attributes, art historical information, visual elements, color, location, description, abstract, content/story, viewer response, and external relationship—were measured. It is interesting to note that *literal object* accounted for up to thirty-one of the responses. Human form and other human characteristics accounted for approximately fifteen responses. Color, texture, and so on ranked much lower compared to the first two categories. The role of content/story varied widely from insignificant to highly important. In other words, users dynamically combine image content and context in their queries.

Romer (1993) describes a wish list for image archive managers, specifically the types of data descriptions necessary for practical retrieval. The heavy reliance on text-based descriptions is questioned. Furthermore, the adaptation of such techniques to multimodal content is required. The need for visual thesauri (Srihari & Burhans, 1994; Chang & Lee, 1991) is also stressed, since these provide a natural way of cataloging pictures, an important task. An ontology of picture types would be desirable. Finally, Romer (1995) describes the need for "a precise definition of image elements and their proximal relationship to one another." This would permit queries such as *Find a man sitting in a carriage in front of Niagara Falls.*

Based on the above analysis, it is clear that object recognition is a highly desirable component of picture description. Although object recognition in general is not possible, for specific classes of objects, and with feedback from text processing, object recognition may be attempted. It is also necessary to extract further semantic attributes of a picture by mapping low-level image features such as color and texture into semantic primitives. Efforts in this area (see Smith & Chang, 1996) are a start but suffer from weak disambiguation and hence can be applied in select databases; our work aims to improve this. Improved text-based techniques for predicting image elements and their structural relationships are presented.

Webpic: A Multimodal Picture Retrieval System

To demonstrate the effectiveness of combining text and image content, a robust, efficient, and sophisticated picture search engine has been developed; specifically, Webpic will selectively retrieve pictures of people in various contexts. A sample query could be *Find outdoor pictures of Bill Clinton with Hillary talking to reporters on Martha's Vineyard.* This should generate pictures where (1) Bill and Hillary Clinton actually appear in the picture (verified by face detection/recognition), and (2) the collateral

text supports the additional contextual requirements. The word "robust" means the ability to perform under various data conditions; potential problems could be lack of, or limited, accompanying text/HTML, complex document layout, and so on. The system should degrade gracefully under such conditions. Efficiency refers primarily to the time required for retrievals which are performed online. Since image indexing operations are time-consuming, they are performed offline. Finally, sophistication refers to the specificity of the query/response. In order to provide adequate responses to specific queries, it is necessary to perform more complex indexing of these data.

Figure 2 depicts the overall structure of the system. It consists of three phases. Phase 1 is the data acquisition phase—multimodal documents from WWW news sites (e.g., MSNBC, CNN, USA Today) are downloaded. In order to control the quality of data that are initially downloaded, a Web crawler in Java has been implemented to do more extensive filtering of both text and images.

The inputs to the system are a set of name keys (names of people) and an initial set of URLs to initiate the search. Some preprocessing tools are employed during this phase. One such tool is an image-based *photograph versus graphic* filter. This filter is designed and implemented based on histogram analysis. Presumably, a photograph histogram has a much wider spectrum than that of a graphic image.

A *collateral text extractor*, whose task is to determine the scope of text relevant to a given picture, is also employed. Caption text appears in a wide variety of styles. News sites such as CNN and MSNBC use *explicit* captions for pictures. These are indicated through the use of special fonts and careful placement using HTML commands as illustrated in Figure 1. In other Web pages, captions are not set off explicitly but, rather, are *implicit* by virtue of their proximity to the picture.

Explicit captions are detected based on the presence of strong HTML clues as well as the usage of key phrases such as "left, foreground, rear" and so on. These can be used to predict picture contents. General collateral text is detected based on the presence of words from the "ALT" tag, caption words, spatial proximity to picture, and so on. Such text, while not a powerful predictor of the contents of a picture, establishes the context of a picture. An image-based caption extractor that extracts ASCII text that has been embedded in images (a common practice among news oriented sites) has been developed in our laboratory and is available for use.

Phase 2 is the content analysis or indexing phase (performed offline). Phase 2 illustrates that both natural language processing (NLP) and image processing result in factual assertions to the database. This represents a more semantic analysis of the data than general text and image indexing based on statistical features. This is discussed in later sections.

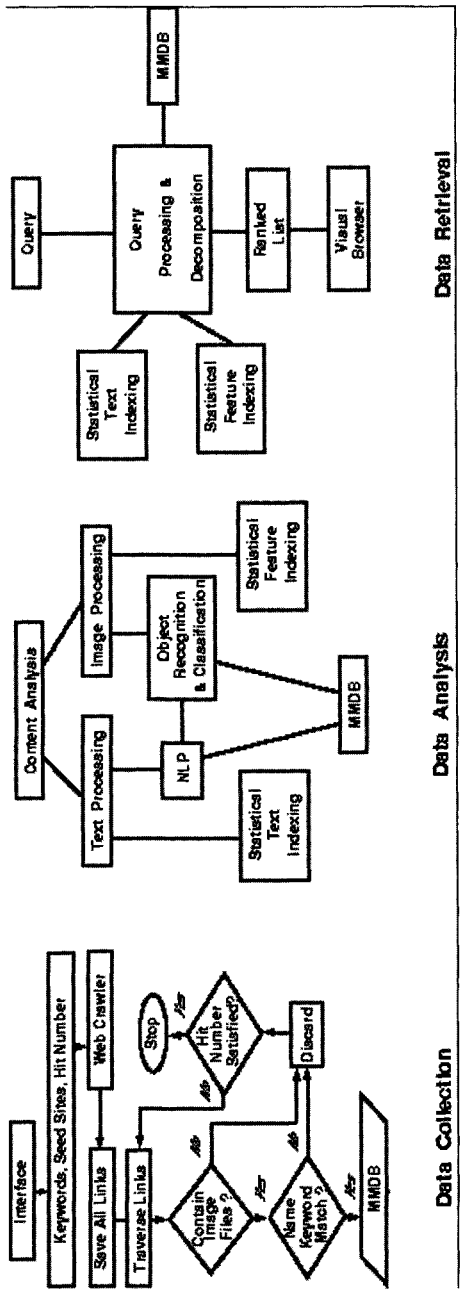


Figure 2. Overall Control Structure of the Proposed System.

Phase 3, retrieval, demonstrates the need to decompose the query into its constituent parts. A Web-based graphical user interface (GUI) has been developed for this. As Figure 3 illustrates, the system permits users to view the results of a match based on different visual criteria. This is especially useful in cases where the user knows the general context of the picture but would like to interactively browse and select pictures containing his or her desired visual attributes. The interface also illustrates that further query refinement using techniques such as image similarity are possible. Finally, although the example illustrates a primary context query, it is possible for the original query to be based on pure image matching techniques. The basic database infrastucture for a multimodal database has been built using Illustra (Illustra is a relational database management system from Informix Inc.).

This is used for data storage as well as representing factual (exact) information. Illustra's ability to define new data types and associated indexing and matching functions is useful for this project.

Metadata

For each picture and its accompanying text, the following metadata are extracted and stored. The metadata model described here is currently applicable only to text and image sources. However, it can be easily extended to accommodate audio and video sources as well:

- Text_Indx: text index, using statistical vector-space indexing techniques. This is useful in judging similarity of two contexts.
- Img_Indx1,Img_Indx2,...Img_Indxk: indexes for various image features based on statistical techniques. This includes color, texture, shape, as well as other properties useful in judging the similarity of two images.
- PDT: this is a template containing information about people, objects, events, locations, and dates mentioned in the text accompanying a picture. Such information is extracted through NLP techniques and will be discussed in the text processing section. Similarity of these templates involves a sophisticated *unification* algorithm.
- Objects: this is a template containing information about objects detected in the image (image coordinates) and their spatial relationships. It also includes information pertaining to general scene classification (e.g., indoor/outdoor, man-made/natural, and so on).

Text Indexing

Text Processing

The goal of natural language processing research in this project is to examine the use of language patterns in collateral text to indicate scene

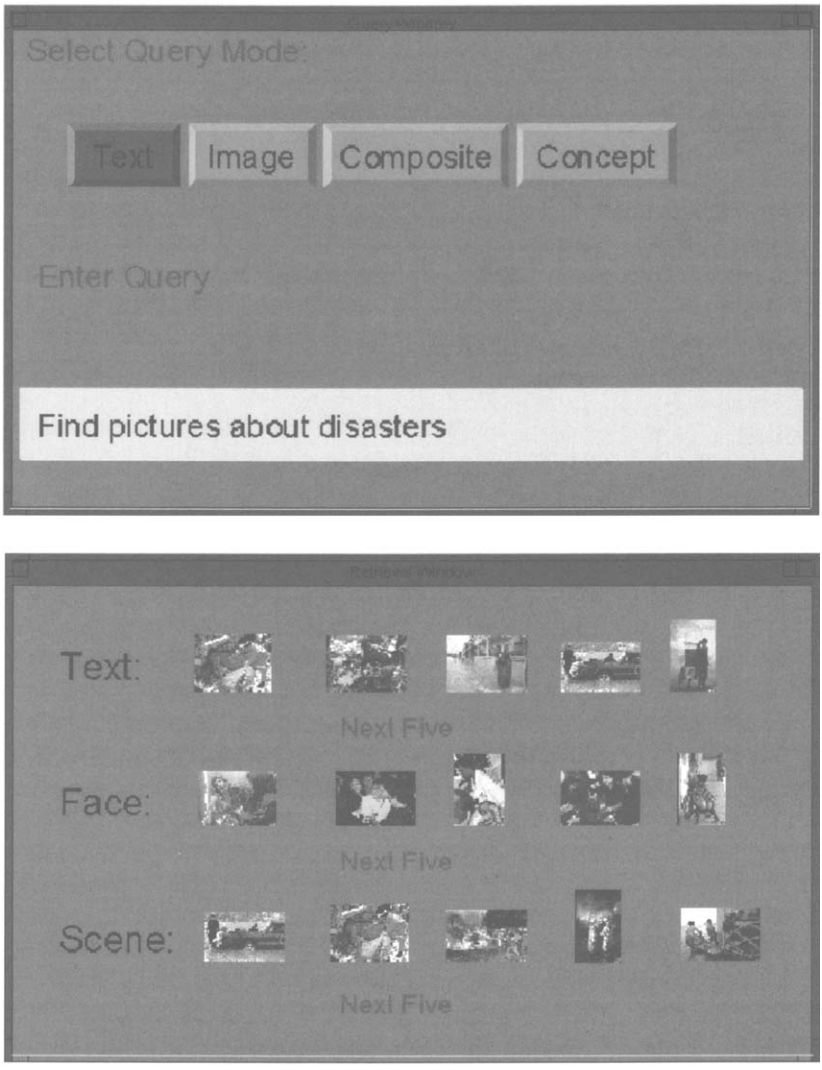


Figure 3. The Multimodal GUI Used in Retrieval.

contents in an accompanying picture. In this section, NLP techniques to achieve this goal are described. The objective is to extract properties of the accompanying picture as well as cataloging the context in which the picture appeared. Specifically, the interest is in deriving the following information that photo archivists have deemed to be important in picture retrieval:

- Determining which objects and people are present in the scene; the location and time are also of importance, as is the focus of the picture.
- Preserving event (or activity) as well as spatial relationships that are mentioned in the text. Spatial information, when present, can be used for automatically identifying people in pictures.

  Consider the caption *President Clinton and his family visited Niagara Falls yesterday. The First Lady and Chelsea went for a ride on the Maid of the Mist.* This should not match the query *find pictures of Clinton on the Maid of the Mist.* However, the caption *Clinton rode the Maid of the Mist Sunday* should be returned. Current IR systems that rely on statistical processing would return both captions. NLP techniques are required for correct processing in this case.
- Determining further attributes of the picture such as indoor versus outdoor, mood, and so on.
- Representing and classifying the general context indicated by the text—e.g., political, entertainment, and so on.

Some organizations, such as Kodak, are manually annotating picture and video clip databases to permit flexible retrieval. Annotation consists of adding logical assertions regarding important entities and relationships in a picture. These are then used in an expert system for retrieval. Aslandogan et al. (1997) describe a system for image retrieval based on matching manually entered entities and attributes of pictures, whereas our objective is to *automatically extract* as much information as possible from natural language captions.

Specifically, the goal is to complete *picture description templates* (PDT) which represent image characteristics. Templates of this type are used by photo repository systems, such as the Kodak Picture Exchange (Romer, 1993). The templates carry information about people, objects, relationships, location, as well as other image properties. These properties include: (1) indoor versus outdoor setting, (2) active versus passive scene—i.e., an action shot versus a posed photo, (3) individual versus crowd scene, (4) daytime versus night-time, and (5) mood.

As an example, consider Figure 4 which shows the output template from processing the caption *A woman adds to the floral tribute to Princess Diana outside the gates of Kensington Palace* of Figure 1. Information extraction (IE) techniques (Sundheim, 1995), particularly shallow techniques, can be used effectively for this purpose. Unlike text understanding sys-

People: person (female, PER1)
Objects: flowers
Activity: pay_tribute (PER1, Princess Diana)
Location: Kensington Palace, "outdoor"
Event Date: Monday, Sept. 2, 1997
Focus: PER1

Figure 4. Picture Description Template (PDT).

tems, IE is concerned only with extracting relevant data that have been specified a priori using fixed templates. Such is the situation here.

Specific techniques for deriving the above information are now presented. The techniques fall into three general categories: statistical text indexing, light parsing, and extracting picture attributes.

Statistical Text Indexing

The goal here is to capture the general context represented by collateral text. Though not useful in deriving exact picture descriptions, statistical text indexing plays a key role in a robust multimodal information retrieval system. There has been considerable research in the area of document indexing and retrieval, particularly the vector space indexing techniques (Salton, 1989). The problem being faced here differs from traditional document matching since the text being indexed—viz, collateral text—is frequently very sparse. Minor adjustments are made to existing techniques in order to overcome the sparseness problem. This includes: (1) the use of word triggers (computed from a large corpus) to expand each content word into a set of semantically similar words, and (2) the use of natural language pre-processing in conjunction with statistical indexing. Word triggers refer to the frequent co-occurrence of certain word pairs in a given window size of text (e.g., fifty words). Natural language pre-processing refers to methods, such as Named Entity Tagging (described below), which classify groups of words as person name, location, and so on. While the use of NLP in document indexing and retrieval has met with limited success, the brevity of collateral text calls for more advanced processing.

Light Parsing: Extracting Patterns of Interest

The previous section described general content indexing; these techniques are based on statistics of word, word-pair frequencies, and so on. In this subtask, the focus is on more in-depth syntactic processing of the relevant text; this is treated as an information extraction task. Such systems consist of several hierarchical layers, each of which attempts to extract more specific information from unformatted text.

In the case of photographs, *template entities* are the objects and people appearing in the photograph, *template relationships* include spatial relationships between objects/people, as well as event/activity information. The first layer consists of *named entity tagging*; this is an extremely useful pre-processing technique and has been the subject of considerable research.

Named entity (NE) tagging refers to the process of grouping words and classifying these groups as person name, organization name, place, date, and so on. For example, in the phrase, *Tiger Woods at the River Oaks Club,* River Oaks Club would be classified as a location. Applying NE tagging to collateral text reduces errors typically associated with words having multiple uses. For example, a query to "Find pictures of oaks along a river" should not retrieve the above caption since *River Oaks Club* is tagged as a location. Bikel et al. (1997) describe a statistical method for NE tagging; given a manually truthed corpus of captions and collateral text, it is straightforward to develop an NE tagger. At this point, a rule-based system for NE tagging has been implemented which is giving better than 90 percent accuracy performance.

The next layers of the hierarchical grammar are used for recognizing domain-independent syntactic structures such as noun and verb groupings (assuming that named entity tagging has already taken place); this leads to identification of template entities and basic relationships (i.e., SVO structure). The processing in these layers is confined to the bounds of single sentences. The final layer is where intersentential information is correlated, thus leading to merging of templates. It is here that the final decision on entries in the picture description template are made. For example, one sentence in a caption may refer to Princess Diana seen at her country estate, while a later sentence may refer to the fact that the estate is located outside the village of Althorp, England. In such a situation, template merging would result in the information that, in the specified picture, the location is Althorp, England. This is a form of co-reference that is being exploited. The template also includes general characteristics of the picture which may be detected from either the caption or collateral text. This is discussed in the next section.

The demands for efficient and robust natural language processing systems have caused researchers to investigate alternate formalisms for language modeling. Current information extraction requirements call for the processing of up to 80 MB of text per hour. Researchers have increasingly turned to finite-state processing techniques (Roche & Schabes, 1997). Roche (1997) says that "for the problem of parsing natural language sentences, finite-state models are both efficient and very accurate even in complex linguistic situations" (p. 241). A finite state transducer (FST) is a special case of a finite state automaton (FSA) in which each arc is labeled by a pair of symbols (input and output)

rather than a single symbol. A rule compiler (Kartutnen & Beesley, 1992) takes *regular relations* as input and constructs the corresponding FST. Operations supported by FST that are useful in grammar construction are union, intersection and, particularly, composition. Domain-specific pattern rules (to extract special attributes for a select domain) can be written as a new FST; this new FST can easily be composed with the base system. Hobbs et al. (1997) employs a cascaded set of FSTs to implement a hierarchical grammar for IE. The picture description grammar is currently being implemented as a cascaded FST.

Extracting Picture Attributes

Once the parsing process has been completed, it is possible to attach further attributes to the picture. This includes attributes such as indoor versus outdoor, mood, and so on. By employing the roles that entities take on in the picture description templates, as well as referring to ontologies and gazetteers, it is possible, in some cases, to extract further attributes. For example, if a caption refers to *Clinton on the White House lawn*, it is characterized as an outdoor picture. This is essentially a *unification* process between location types. Chakravarthy (1994) discusses the use of WordNet in performing such characterization.

Image Indexing

Imagery is probably the most frequently encountered modality, next to text, in multimedia information retrieval. Most of the existing techniques in the literature of *content-based retrieval* or *image indexing and retrieval* use low-level or intermediate-level image features such as color, texture, shape, and/or motion for indexing and retrieval. Although these methods may be efficient in retrieval, the retrieval precision may not be good enough, as typically it may not be true that image features always reflect their semantic contents.

In this article, the focus is mainly on image retrieval of people or scenes in a general context. This requires capabilities of face detection and/or recognition in the general image domain. By a general image domain, it is meant that the appearances of the objects in question (e.g., faces) in different images may vary in size, pose, orientation, expression, background, as well as contrast. Since color images are very popular in use and very easy to obtain, these have been chosen for experimentation.

The potential applications of the capability of face detection and/or face recognition include: (1) filtering—i.e., determining whether or not a particular image contains a human being, (2) identifying individuals—i.e., handling queries for certain well-known people using face recognition, and (3) improving the accuracy of similarity matching. For images involving human faces, it is very difficult to check similarity based on

histograms of the entire images. Color histogram techniques do not work well for images containing faces. However, after applying face detection to the original images, the face areas may be automatically "cropped" out, and the rest of the image may still be used for histogram-based similarity matching.

Face detection and/or recognition has received focused attention in the literature of computer vision and pattern recognition for years. A good survey on this topic may be found in Chellappa et al. (1995). Typically, face detection and recognition are treated separately in the literature, and the solutions proposed are normally independent of each other. In this task, a *streamlined solution* to both face detection and face recognition is pursued. By a streamlined solution, it is meant that both detection and recognition are conducted in the same color feature space, and the output of the detection stage is directly fed into the input of the recognition stage. Another major difference between the present research and work described earlier in the literature is that the proposed system is a self-learning system, meaning that the face library used in face recognition is obtained through face detection and text understanding using the earlier research system PICTION (Srihari, 1995b). This allows the stage of face data collection for construction of the face library as an automatic part of data mining, as opposed to interactive manual data collection usually conducted for face recognition. Note that in many situations it is impossible to do manual data collection for certain individuals, such as Bill Clinton. For those people, their face samples can only be obtained through the WWW, newspapers, and so on. Thus, automatic data collection is not only efficient but is also necessary.

Face detection is approached as pattern classification in a color feature space. The detection process is accomplished in two major steps: feature classification and candidate generation. In the feature classification stage, each pixel is classified as face or nonface based on a standard Bayesian rule (Fukunaga, 1990). The classification is conducted based on pre-tuned regions for the human face in a color feature space. The color features used in this approach are hue and chrominance. The pre-tuning of the classification region in the color feature space is conducted by sampling over 100 faces of different races from different Web sites. In the candidate generation stage, first a morphological operation is applied to remove the noise, and then a connected component search is used to collect all the "clusters" that indicate the existence of human faces. Since the pre-tuned color feature region may also classify other parts of the human body as candidates, let alone certain other objects that may happen to be within the region in the color feature space, heuristic checking is used to verify the shape of the returned bounding box to see if it conforms to the "golden ratio" law.[1] Figure 5 shows the whole process of face detection and recognition for a Web

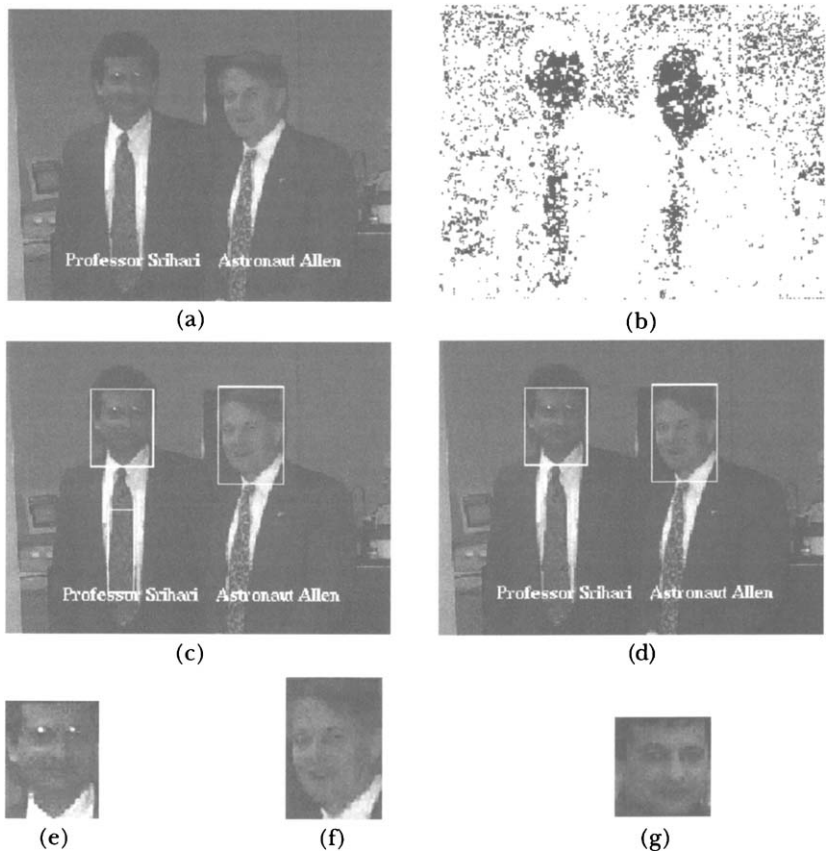


Figure 5. An Example of the Result of Automatic Face Detection and Face Recognition. (a) An original image from internet web. (b) The binary image after classification. (c) Result after morphological operations and connected component search. (d) Final detection result after applying heuristic checking to reject false positives. (e) The first face candidate returned for face recognition. (f) The second face candidate returned for face recognition. (g) Another face image of the same individual as in (e).

image. Note that each detected face is automatically saved into the face library if it has a strong textual indication of who this person is (self-learning to build up the face library), or the face image is searched in the face library to find who the person is, if the query asks to retrieve images of this individual (query stage).

In the face recognition stage, there are two modes of operation. In the mode of face library construction, it is assumed that each face image has its collateral textual information to indicate identities of the people in the image. Face detection is first applied to detect all the faces. Based on

the collateral information (Srihari, 1995b), the identities for each face may be found and thus saved into the library automatically. In the mode of query, on the other hand, the detected face needs to be searched in the library to find out the identity of this individual.

Figure 5(e) and (g) are two face images of the same individual. This is a problem of finding semantic similarity between two face images in the general image domain to identify whether or not two face images contain the same individuals. This is one of the current research directions underway. Promising experimental results based on preliminary tests show that it is possible to include the capability of querying individuals in image retrieval by conducting semantic similarity matching.

To summarize, image processing capability currently consists of: (1) a face detection module based on color feature classification to determine whether or not an image contains human faces, and (2) a histogram-based similarity matching module to determine whether or not two images "look" similar.

Multimodal Query Processing

Even though there has been much success recently in text-based information retrieval systems, there is still a feeling that the needs of users are not being adequately met. Multimodal IR presents an even greater challenge since it adds more data types/modalities, each having its own retrieval models. The body of literature in multimodal IR is vast, ranging from logic formalisms for expressing the syntax and semantics of multimodal queries (Meghini, 1995) to MPEG-4 standards for video coding which call for explicit encoding of semantic scene contents. A popular approach has been to add a layer representing *meta querying* on top of the individual retrieval models. An agent-based architecture for decomposing and processing multimodal queries is discussed in Merialdo and Dubois (1997). In focusing so much on formalisms, especially in the logic-based approaches, researchers sometimes make unreal assumptions about the quality of information that can be automatically extracted (e.g., the detection of complex temporal events in video).

The present research focuses not on the formalism used to represent the queries, rather, the focus is on the effect of utilizing *automatically* extracted information from multimodal data in improved retrieval. Processing queries requires the use of: (1) information generated from statistical text indexing, (2) information generated from natural language processing of text, and (3) information generated from image indexing—in this case, face detection and recognition—as well as color, shape, and texture indexing.

Thus, matching a query to a captioned image in the database could involve four types of similarity computation:

1. *(Text_Indx$_q$, Text_Indx$_{CapImg}$):* text-based similarity, statistical approach;
2. *SIM(Img_Indx(j)$_q$,Img_Indx$_{CapImg}$):* j=1,...,k: image similarity, for each image feature statistical approach;
3. *SIM(PDT$_q$,PDT$_{CapImg}$):* text-based concept similarity, symbolic approach; and
4. *SIM(Objects$_q$,Objects$_{CapImg}$):* image-based content similarity, symbolic approach.

Syntax and Semantics of Multimodal Queries

Similarity matching techniques for each information source are discussed in the next section. Here the discussion centers on the interpretation of the query, as handled by the procedure *Interpret_Query* which attempts to understand the user's request and decompose it accordingly.

User input includes one or more of the following: (1) *text_query*, a text string; (2) *image_query*, an image; (3) *topic_query*, one or more concepts selected from a pre-defined set of topics, such as *sports, politics, entertainment,* and so on; and (4) *user_preferences*, a set of choices made by the user indicating preferred display choices and so on. These are used by the *Interpret_Query* module in determining *ranking schemes.*

The specific objective of the *Interpret_Query* procedure is: (1) to determine the arguments to each of the *SIM(x,y)* components mentioned above, and (2) to determine the set of ranking schemes that will be used in presenting the information to the user. Determining arguments to the text and image similarity functions are straightforward. The text string comprising the query is processed, resulting in content terms to be used in a vector-space matching algorithm. In the case of a query image, the image features are available already, or are computed if necessary. Determining the arguments to the picture description template similarity and object similarity are more involved. Some natural language processing analysis of the *Text_String* is required to determine which people, objects, events, and spatial relationships are implied by the query.

Another important issue is to decide on how information should be combined. For example, for an unambiguous query such as *Find pictures of Bill Clinton,* the face detection and recognition results will be automatically applied to produce a single ranking of images satisfying the query. However, for a more subjective query, such as *Find pictures of victims of natural disasters,* the general context is first applied. The results are then sorted based on various visual criteria, thus allowing the user to browse and make a selection.

Each ranking scheme *(RSk)* defines a ranking (CapImg(k,1),CapImg (k,2),...,CapImg(k,nk)) of the images in the database. Currently, a simple technique to generate ranking schemes is employed. For each information source that is involved in a query, several sort criteria are applied in varying order. These sort criteria reflect the relative importance of each

information source. For example, for queries involving finding people in various contexts, two sorted lists will be presented to the user. The first weights the context more and the second weights the face detection results more—i.e., presence of face, relative size of face.

Matching Queries to Data

Text-based similarity is based on statistical indexing techniques; while not as precise as natural language processing techniques, it is very robust. Image-based similarity techniques using color, shape, texture, and so on have been discussed extensively in the content-based image retrieval literature. Image-based content similarity includes any visual information that has been verified by using object recognition techniques (e.g., number of faces, gender) or semantic classification (e.g., indoor versus outdoor).

When matching based on the similarity of picture description templates, it is necessary to employ unification techniques. For example, a search for *Dalmation* should match a picture whose PDT contains *dog*. That is, *Unify(Dalmation,dog)* should return a non-zero value. An approach similar to that of Aslandogan et al. (1997) to perform inexact matching is being adopted. The use of ontologies is required for several purposes in this phase. First, they are required to map entities into their *basic categories* (Rosch et al., 1976); research has shown that people most often query by basic categories (e.g., *dog* rather than *Doberman*). If the caption refers to the location as an *auditorium*, for example, it is necessary to map this into *building* for the purpose of retrieval. Similar mapping needs to take place on query terms. Srihari (1995a) and Aslandogan et al. (1997) discuss the use of WordNet in matching picture entities with queries. WordNet provides critical information in determining hierarchical relationships between entity classes and event classes.

Query Refinement and Relevance Feedback

Since users are not always sure of what they are looking for, an adaptive system is required. After specifying an initial query, the results are sorted into various classes based on the ranking schemes suggested by *Interpret_Query*. Users may choose to refine the query by either modifying the text query, concept query, or select images that best match their needs. The latter are used in a relevance feedback process, where users can interactively select pictures that satisfy their needs. Although the technique is well-understood in the text domain (Chang, 1998; Robertson, 1986; Rocchio, 1971; Ide, 1971; Croft & Harper, 1979; Fuhr & Buckley, 1991), it is still in the experimental stage in the image domain (Smith, 1997). Popular techniques include Rocchio's (1971) relevance feedback formula for the vector model and its variations (Ide, 1971), and the Croft-Harper formula (1979) for the probabilistic retrieval model and its modifications (Fuhr & Buckley, 1991; Robertson, 1986). Query refinement consists of

adjusting the weights assigned to each feature; this is the technique adopted in the text domain. Of course, the difficult aspect is determining which features are important. The multiple ranking scheme described in the previous section is of use here since each ranking corresponds to the importance of certain features (or metadata). By selecting images in certain ranking schemes, the system is able to learn which features are useful. This process can continue iteratively until the user finds the required picture. The user interface supports the visual browsing that is an integral part of image retrieval.

Retrieval Experiments

There were two experiments conducted in picture retrieval from multimodal documents. Each reflected a different strategy of combining information obtained by text indexing and image indexing. Both of these are now described.

Single Ranking Method

In this experiment, the queries are first processed using text indexing methods. This produces a ranking *Px1, ... , Pxn* as indicated in Figure 6.

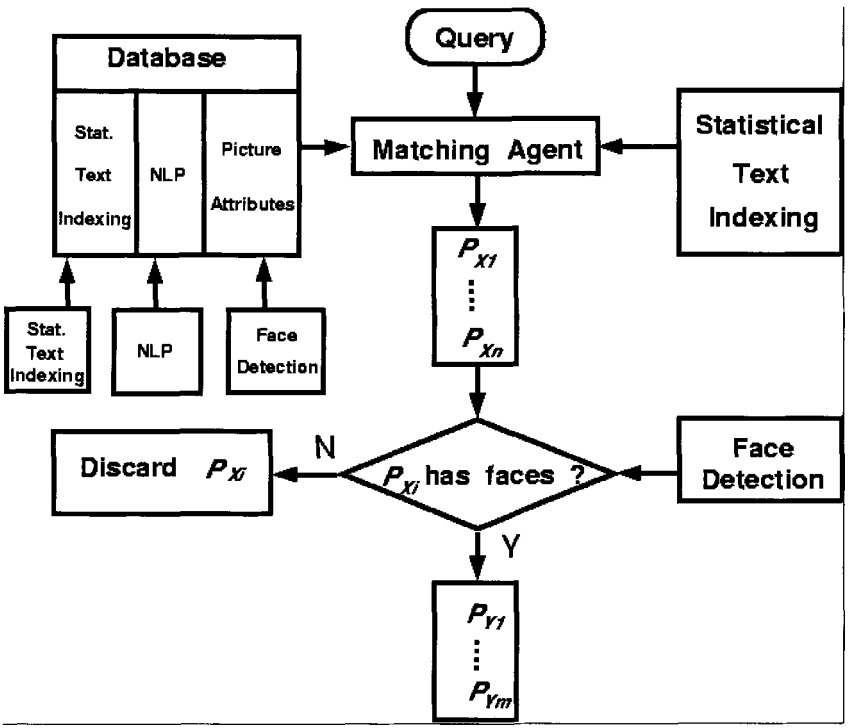


Figure 6. Single-Ranking Strategy for Combining Text and Image Information in Retrieval.

Those pictures *pxi* which do not contain faces are subsequently eliminated; this information is obtained by running the automatic face detection module.

Figure 7 presents the results of an experiment conducted on 198 images that were downloaded from various news Web sites. The original data set consisted of 941 images. Of these, 117 were empty (white space) and 277 were discarded as being graphics. From these, a subset of 198 images was chosen for this experiment.

	Text Only	Text + Manual Insp	Text + Face Det
At 5 docs	1.0	1.0	1.0(3)
At 10 docs	1.0	0.70	1.0(3)
At 15 docs	0.80	0.75	1.0(2)
At 30 docs	0.77	0.67	NA

Figure 7. Results of Single Ranking Strategy of Combining Text and Image Content. The last column indicates the result of text indexing combined with face detection. The number in parentheses indicates the number of images in the given quantile that were discarded due to failure to detect faces.

There were ten queries, each involving the search for pictures of named individual(s); some specified contexts also, such as *find pictures of Hillary Clinton at the Democratic Convention.* Due to the demands of truthing, the results for one query are reported; more comprehensive evaluation is currently underway. Figure 7 indicates precision rates using various criteria for content verification: (1) using text indexing (SMART) alone, (2) using text indexing and manual visual inspection, and (3) using text indexing and automatic face identification. As the table indicates, using text alone can be misleading—when inspected, many of the pictures do not contain the specified face. By applying face detection to the result of text indexing, photographs that do not have a high likelihood of containing faces are discarded. The last column indicates that this strategy is effective in increasing precision rates. The number in parentheses indicates the number of images in the qiven quantile that were discarded due to failure to detect faces.

Sample output is shown in Figures 8 and 9. Figure 8 illustrates the output based on text indexing alone. The last picture illustrates that text alone can be misleading. Figure 9 illustrates the re-ranked output based on results from face detection. This has the desired result that the top images are all relevant. However, a careful examination reveals that, due to the face detector's occasional failure to detect faces in images, relevant images are inadvertently being discarded. Thus this technique increases precision but lowers recall. However, if the source of images is the WWW, this may not be of concern. The face detector is continually being improved to make it more robust to varied lighting conditions.

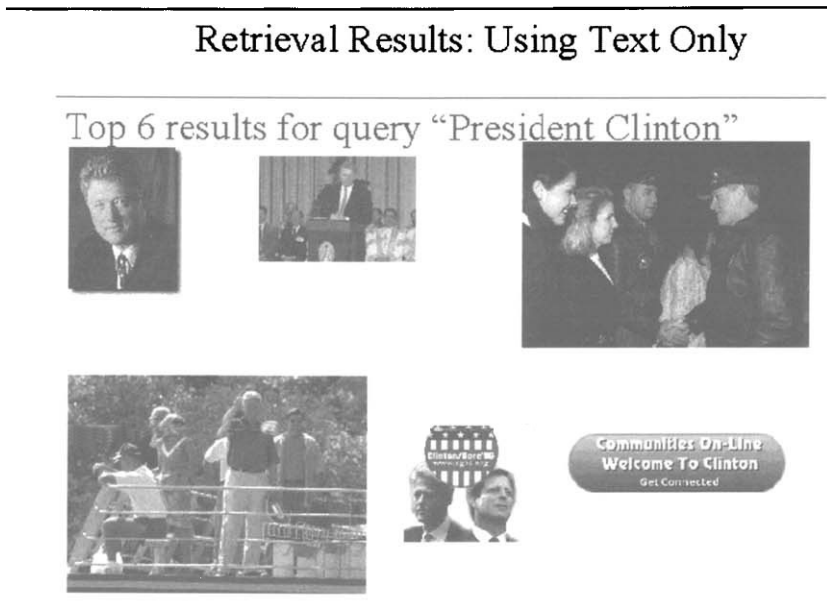


Figure 8. Top Six Images Based on Text Indexing Only.

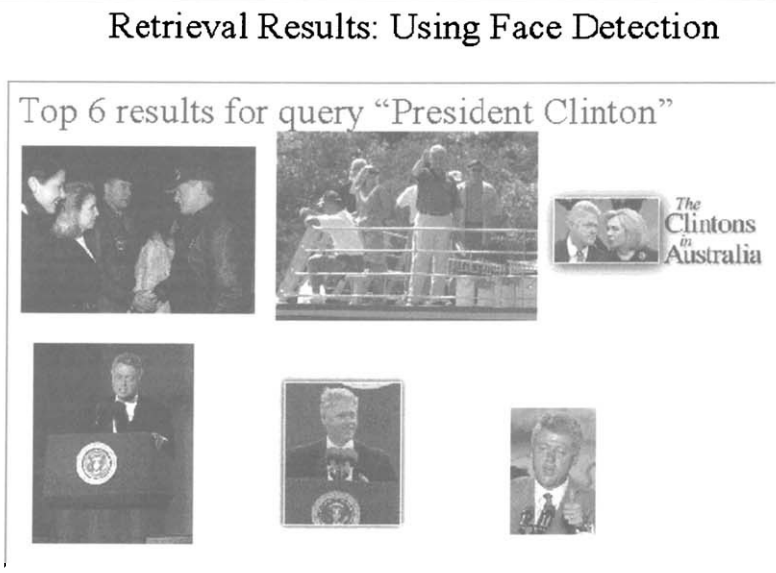


Figure 9. Top Six Images Based on Combining Text Indexing and Face Detection.

Multiple Ranking Method

In this experiment, a multiple ranking method for presenting candidate images to the user is employed. This strategy is depicted in Figure 10. The context is first verified using statistical text indexing. These candidate images are then sorted based on various visual properties. The first property is the presence of faces, the second represents the absence of faces (reflecting an emphasis on general scene context rather than individuals). This reflects the assumption that users do not know a priori exactly what kind of pictorial attributes they are looking for—i.e., that they would like to browse. Figure 11 depicts the top ranked images for the query *victims of disasters.*

Many of these refer to the recent air crash in Indonesia, partially blamed on heavy smoke from forest fires. Some images depict victims, some depict politicians discussing the situation. Based on an imposed threshold, only the top ten images returned by text retrieval were considered. As the results show, this produces a different ranking of images, where the lower row clearly emphasizes people. Had a lower threshold for text retrieval been used, the difference would have been more dramatic.

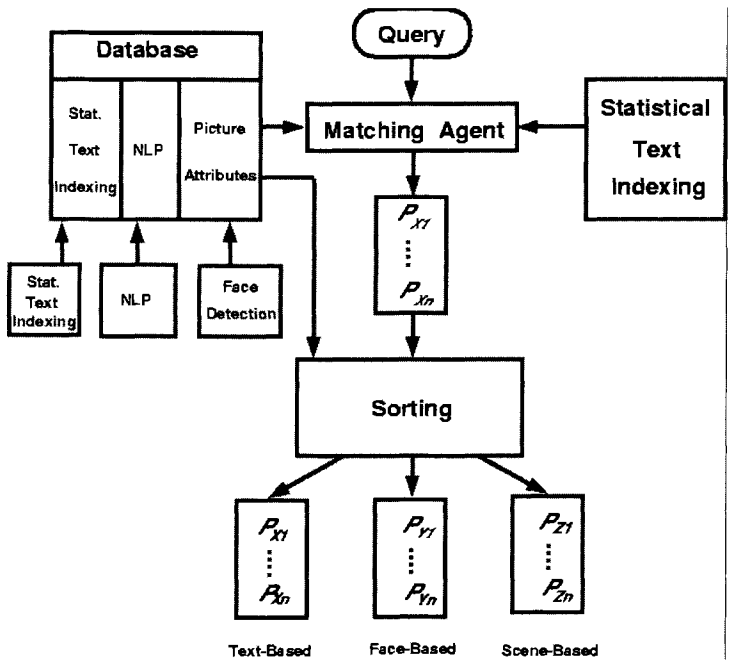


Figure 10. Multiple Ranking Strategy for Combining Text and Image Information in Retrieval.

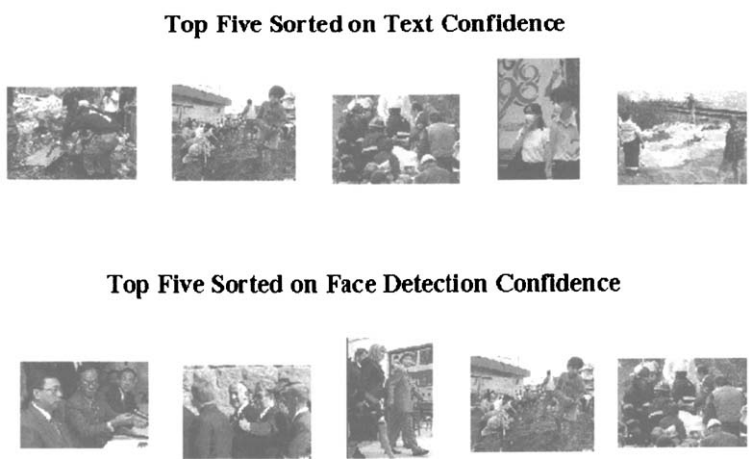


Figure 11. Top Five Images Based on Combining Text Indexing and Face Detection.

Evaluating precision and recall for such a technique is challenging. The precision rate for a given sorting criterion is based on both the text relevance and the presence of the required pictorial attributes (e.g., presence of faces). The text retrieval precision for the top ten images is 90 percent. However, when "presence of faces" is used as a sorting criterion, the precision in the top ten images drops to 40 percent. This is primarily due to the presence of very small faces in the image which are found by the face detector. Since the manual annotators were instructed to disregard faces below a certain size, these are judged to be erroneous (e.g., the last picture in the second row of Figure 11). Thus, assigning relevance judgments based on pictorial attributes must be reinvestigated.

Future Directions

Future directions include improvements on several fronts. First, it is necessary to incorporate information derived from natural language processing as well as statistical image indexing into the retrieval model. Second, the experiments conducted so far have involved only a single query modality, namely text. The next step is to permit multimodal queries, whereby the user can specify an information request using a combination of text (representing contextual constraints) and images (representing exemplars). A relevance feedback mechanism whereby the system can "learn" from user feedback is called for.

Finally, there is a need for more comprehensive testing and evaluation of the techniques developed thus far. The development of evaluation frameworks suitable for multimedia information retrieval systems is still an emerging research area. It is the focus of the MIRA (1999) project,

a consortium of IR researchers in Europe. They make a strong case for *dynamic evaluation* techniques for such applications as opposed to the static evaluation techniques used in text retrieval systems. Rather than evaluating the initial results of a single query, researchers are proposing that the evaluation should be associated with an entire session consisting of continuously refined queries. For example, a monotonically increasing performance curve indicates a good session. They also suggest that new interaction-oriented tasks (apart from search and retrieval) must be supported and evaluated. An example of the latter would be the ability to clarify and formulate information needs.

In this research effort, the following measures of performance are of interest: (1)effectiveness of the ranking scheme generated based on the user's query input and preferences, (2) performance of each individual ranking scheme, and (3) performance of the face detection and recognition modules.

Conclusion

This article has presented a system for searching multimodal documents for pictures in context. Several techniques for extracting metadata from both images and text have been introduced. Two different techniques for combining information from text processing and image processing in the retrieval stage have been presented. This work represents efforts toward satisfying users' needs to browse efficiently for pictures. It is also one of the first efforts to automatically derive semantic attributes of a picture, and to subsequently use this in content-based retrieval. Retrieval experiments discussed in this article have utilized only two of the four indexing schemes that have been developed. These show the promise of integrating several modalities in both the indexing and retrieval stages.

Notes

[1] It is believed that for a typical human face, the ratio of the width to the height of the face is always around the magic value of 2/(1+05), which is called the *golden ratio* (Farkas & Munro, 1987).

References

Aslandogan, Y. A.; Their, C.; Yu, C.T.; Zou, J.; & Rishe, N. (1997). Using semantic contents and WordNet in image retrieval. In *SIGIR '97* (Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 27-31, 1997, Philadelphia, PA) (pp. 286-295). New York: Association for Computing Machinery.

Chang, C. C., & Lee, S. Y. (1991). Retrieval of similar pictures on pictorial databases. *PatternRecognition, 24*(7), 675-680.

Chang, W. C. (1998). *A framework for global integration of distributed visual information systems.* Unpublished doctoral dissertation, State University of New York, Buffalo.

Charkravarthy, A. S. (1994). Representing information need with semantic relations. In *COLING-94* (The 15th International Conference on Computational Linguistics, August 5-9, 1994, Kyoto, Japan) (pp. 737-741). Morristown, NJ: ACL.

Chellappa, R.; Wilson, C.; & Sirohey, S. (1995). Human and machine recognition of faces: A survey. *Proceedings of the IEEE, 83*(5), 705-741.

Croft, W., & Harper, D. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of Documentation, 35*(4), 285-295.

Farkas, L. G., & Munro, I. R. (1987). *Anthropometric facial proportions in medicine.* Springfield, IL: Charles C. Thomas.

Fuhr, N., & Buckley, C. (1991). A probabilistic learning approach for document indexing. *ACM Transactions on Information Systems, 9*(3), 223-248.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition* (2d ed.). Boston: Academic Press.

Hobbs, J. R.; Appelt, D.; Bear, J.; Israel, D.; Kameyama, M.; Stickel, M.; Tyson, M. (1997). Fastus: A cascaded finite-state transducer for extracting information from natural language text. In E. Roche & Y. Schabes (Eds.), *Finite-state language processing* (pp. 383-406). Cambridge, MA: MIT.

Ide, E. (1971). New experiments in relevance feedback. In G. Salton (Ed.), *The SMART system: Experiments in automatic document processing* (pp. 337-354). Englewood Cliffs, NJ:Prentice Hall.

Jorgensen, C. (1996). An investigation of pictorial image attributes in descriptive tasks. In B. E. Rogowitz & J. P. Allenbach (Eds.), *Human vision and electronic imaging* (Proceedings of the Society for Optical Engineering (vol. 2657, pp. 241-251). Bellingham, WA: SPIE.

Kartutnen, L., & Beesley, K. R. (1992). *Two-level rule compiler.* Unpublished Xerox PARC Tech. Report No. TR ISTL-92-2.

Meghini, C. (1995). An image retrieval model based on classical logic. In *SIGIR '95* (Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 9-13, 1995, Seattle, WA) (pp. 300-309). New York: Association for Computing Machinery Press.

Merialdo, B., & Dubois, F. (1997). An agent-based architecture for content-based multimedia browsing. In M. T. Maybury (Ed.), *Intelligent multimedia information retrieval* (pp. 281-294). Cambridge, MA: AAAI Press.

MIRA. (1999). *Evaluation frameworks for interactive multimedia information retrieval applications.* Retrieved July 7, 1999 from the World Wide Web: http://www.dcs.gla.ac.uk/mira.

Robertson, S. (1986). On relevance weight estimation and query expansion. *Journal of Documentation, 42*(3), 182-188.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.), *The SMART retrieval system: Experiments in automatic document processing* (pp. 313-323). Englewood Cliffs, NJ: Prentice-Hall.

Roche, E. (1997). Parsing with finite-state transducers. In E. Roche & Y. Schabes (Eds.), *Finite-state language processing* (pp. 241-280). Cambridge, MA: MIT.

Romer, D. M. (1993). *A keyword is worth 1,000 images* (Kodak Internal Tech. Rep.). Rochester, NY: Eastman Kodak.

Romer, D. M. (1995). *Research agenda for cultural heritage on information networks.* Retrieved July 7, 1999 from the World Wide Web: http://www.ahip.getty.edu/agenda.

Rosch, E.; Mervis, C. B.; Gray, W. D.; Johnson, D. M.; Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*(3), 382-439.

Rowe, N., & Guglielmo, E. (1993). Exploiting captions in retrieval of multimedia data. *Information Processing and Management, 29*(4), 453-461.

Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer.* Reading, MA: Addison-Wesley.

Smith, J. R. (1997). *Integrated spatial and feature image systems: Retrieval, analysis, and compression.* Unpublished doctoral dissertation, Columbia University, New York.

Smith, J. R., & Chang, S.-F. (1996). VisualSEEK: A fully automated content-based image query system. In *Proceedings of ACM Multimedia '96* (November 18-22, 1996, Boston, MA) (pp. 87-98). New York: Association for Computing Machinery Press.

Srihari, R. K. (1995a). Automatic indexing and content-based retrieval of captioned images. *Computer, 28*(9), 49-56.

Srihari, R. K. (1995b). Use of captions and other collateral text in understanding photographs. *Artificial Intelligence Review, 8*(5-6), 409-430.

Srihari, R. K., & Burhans, D. T. (1994). Visual semantics: Extracting visual information from text accompanying pictures. In *Proceedings of the Twelfth National Conference on Artificial Intelligence* (pp. 793-798). Menlo Park, CA: AAAI Press.

Sundheim, B. (Ed.). (1995). *MUC-6* (Proceedings of the 6th Message Understanding Conference, November 6-8, 1995, Columbia, MD). San Francisco: Morgan Kaufmann.

Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision, 7*(1), 11-32.

Additional Reference

Bikel, D. M.; Miller, S.; Schwartz, R.; & Weischedel, R. (1997). Nymble: A high-performance learning name-finder. In *Proceedings of the 5th Conference on Applied Natural Language Processing* (March 31-April 3 1997, Washington DC) (pp. 194-201). Boston: MIT Press.

About the Contributors

CAROLINE R. ARMS is a Program Coordinator for the National Digital Library Program at the Library of Congress. She is responsible for the integration into American Memory of collections digitized at other institutions through the Library of Congress/Ameritech National Digital Library Competition. In the late 1980s, she edited two books of case studies for EDUCOM—*Campus Networking Strategies,* and *Campus Strategies for Libraries and Electronic Information.* More recent publications include "Historical Collections for the National Digital Library" (*D-Lib Magazine,* April & May 1996) and *Enabling Access in Digital Libraries: A Report on a Workshop on Access Management* (Digital Library Federation, 1999).

THERESA GROSE BEAMSLEY is the Director of Collections Information Resources at the Henry Ford Museum & Greenfield Village. This unit includes the Library, Research Center, Archives, Collections Information Management, and Information Delivery and Design departments. She is responsible for the technical design and management of all collections-related electronic information services, including the institution's Web site. Ms. Beamsley holds advanced degrees in Social/Cultural Anthropology and Information Science and brings more than fifteen years of software design and development experience to her position. She is a member and frequent speaker at the annual meetings of the American Society for Information Science and numerous other museum and technical professional societies.

HSIN-LIANG CHEN is an Assistant Professor in the School of Library and Information Science at the University of Wisconsin-Milwaukee. His research interests are image retrieval, human-computer interaction, instructional technology, user studies, and information literacy.

D. A. FORSYTH is presently an Associate Professor of Computer Science at the University of California at Berkeley. He was an Assistant Professor of Computer Science at the University of Iowa and an Assistant Professor of Computer Science at the University of California at Berkeley. His interests include object recognition.

SAMANTHA K. HASTINGS is a faculty member in the School of Library and Information Sciences, University of North Texas in Denton. Ms. Hastings teaches a variety of courses including indexing and abstracting and telecommunications. She runs a program of study for digital image managers with the help of a grant from the federal Institute of Museum and Library Services. In addition, the grant funds a study investigating the impact of Web access to the collections at the African American Museum of Art in Dallas, Texas.

P. BRYAN HEIDORN is an instructor and researcher at the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign where he joined the faculty in 1995. Mr. Heidorn's research interests include natural language processing, spatial cognitive modeling, and image storage and retrieval. His current work involves natural language understanding for the generation of metric models for image synthesis and retrieval. He teaches in the areas of information system automation and information retrieval. Mr. Heidorn is an active member of the American Society of Information Science and the American Society for Computing Machinery.

THOMAS S. HUANG joined the University of Illinois at Urbana-Champaign in 1980, where he is now William L. Everitt Distinguished Professor of Electrical and Computer Engineering, Research Professor at the Coordinated Science Laboratory, and Head of the Image Formation and Processing Group at the Beckman Institute for Advanced Science and Technology. He was on the Faculty of the Department of Electrical Engineering at MIT from 1963 to 1973 and on the faculty of the School of Electrical Engineering and Director of its Laboratory for Information and Signal Processing at Purdue University from 1973 to 1980. Dr. Huang's professional interests lie in the broad area of information technology, especially the transmission and processing of multidimensional signals. He has published twelve books and over 300 papers on network theory, digital filtering, image processing, and computer vision. He received the IEEE Acoustics, Speech, and Signal Processing Society's Technical Achievement Award in 1987 and the Society Award in 1991. He is a Founding Editor of the *International Journal of Computer Vision, Graphics, and Image Processing* and editor of the *Springer Series in Information Sciences* published by Springer Verlag.

SHARAD MEHROTRA is an Assistant Professor in the Computer Science Department at the University of Illinois at Urbana-Champaign since 1994. He has subsequently worked at MITL, Princeton, as a scientist from 1993 to 1994. He specializes in the areas of database management, distributed systems, and information retrieval. His current research projects are on multimedia analysis, content-based retrieval of multimedia objects, multidimensional indexing, uncertainty management in databases, and concurrency and transaction management. Dr. Mehrotra is an author of over fifty research publications in these areas. He is the recipient of the NSF Career Award and the Bill Gear Outstanding junior faculty award in 1997.

MICHAEL ORTEGA is currently pursuing his graduate studies at the University of Illinois at Urbana-Champaign. He received a Fulbright/CONACYT/García Robles scholarship to pursue graduate studies as well as the Mavis Award at the University of Illinois and is a member of the Phi Kappa Phi honor society, the IEEE computer society, and member of the ACM. His research interests include multimedia databases, database optimization for uncertainty support, and content-based multimedia information retrieval.

EDIE RASMUSSEN is an Associate Professor in the School of Information Sciences at the University of Pittsburgh, where she teaches courses in information retrieval, image database management, and indexing and abstracting. She has also taught and researched in institutions in Canada, Malaysia, Singapore, and the United Kingdom.

NEIL C. ROWE is Associate Professor of Computer Science at the U.S. Naval Postgraduate School where he has been since 1983. His main research interest is intelligent access to multimedia databases, and he has also done work on robotic path planning, intelligent tutoring systems, and computer security. Mr. Rowe is the author of *Artificial Intelligence Through Prolog* (Prentice-Hall, 1988) and a forthcoming book on information security.

YONG RUI is currently a researcher at Microsoft Research in Redmond, Washington. His research interests include multimedia information retrieval, multimedia signal processing, computer vision, and artificial intelligence. He has published over thirty technical papers in these areas. He is a 1989-1990 Huitong University Fellowship recipient, a 1992-1993 Guanghua University Fellowship recipient, and a 1996-1998 CSE Engineering College Fellowship recipient.

BETH SANDORE is Head of the Digital Imaging and Multimedia Initiatives program and Associate Professor at the University of Illinois at Urbana-Champaign Library. Her professional experience and research focus

on technology development and evaluation in libraries, including experimental work with image and multimedia databases. Her recent publications include a user evaluation study of the MESL image database, funded by the Getty Information Institute, a book on technology and management in libraries co-authored with F. W. Lancaster, and the proceedings of the 1996 Digital Image Access and Retrieval Conference on experimental digital image database development. Ms. Sandore is active in the American Library Association's Library and Information Technology Association (LITA). She has served in an advisory capacity for a number of groups on imaging and technology evaluation projects, including the U.S. Department of Education, the Getty Information Institute, and the Andrew Mellon Foundation.

ROHINI K. SRIHARI is an Associate Professor of Computer Science and Engineering at SUNY at Buffalo. She has worked in both natural language processing as well as computer vision. Dr. Srihari's current research interests include multimedia information retrieval and multimodal image annotation systems. She is presently working on three projects. The first project, Show&Tell—a multimodal system (combining speech, deictic input, and computer vision) for aerial image annotation and retrieval—was sponsored by the DOD as part of the RADIUS image exploitation program. The second project, WebPiction—for combining text and image context in image retrieval for the World Wide Web and Imagination—is an extension of a DOD sponsored effort on the use of collateral text in image understanding. The third project, Imagination—an Image Annotation and Metadata Generation System for Consumer Photos—is being sponsored by Kodak as part of their digital imaging initiative. Dr. Srihari recently organized a workshop on Multimedia Indexing and Retrieval in conjunction with the ACM conference on Information Retrieval, SIGIR '99.

CHRISTIE STEPHENSON is Librarian for Digital Collections at the New York University Libraries. She served as Project Director of the Museum Educational Site Licensing Project from September 1996 until the conclusion of the project. Prior to that, she served as MESL Project Coordinator and Coordinator of the Digital Image Center at the University of Virginia Library, where she was Assistant Fine Arts Librarian.

ZHONGFEI ZHANG is an Assistant Professor in the Computer Science Department at SUNY at Binghamton. Prior to that he was a Research Assistant Professor at the Department of Computer Science and Engineering at SUNY Buffalo. His research interests include multimedia information indexing and retrieval, image understanding and processing, pattern recognition, artificial intelligence, and robotics. He has published over twenty academic papers in international and national journals and conferences.

STATEMENT OF OWNERSHIP, MANAGEMENT, AND CIRCULATION

(1) Publication Title: *Library Trends.* (2) Publication Number: 0024-2594. (3) Filing Date: October 1999. (4) Issue Frequency: Quarterly (Summer, Fall, Winter, Spring). (5) Number of Issues Published Annually: Four. (6) Annual Subscription Price: $85. (7) Complete Mailing Address of Known Office of Publication (Not printer) (Street, city, county, state, and ZIP+4): University of Illinois, Graduate School of Library and Information Science, Publications Office, 501 East Daniel Street, Champaign, Illinois 61820-6211, Champaign County; Contact Person: Marlo Welshons; Telephone: (217) 333-1359. (8) Complete Mailing Address of Headquarters or General Business Office of Publisher (Not printer): University of Illinois, Graduate School of Library and Information Science, Publications Office, 501 East Daniel Street, Champaign, Illinois 61820-6211. (9) Full Names and Complete Mailing Addresses of Publisher, Editor, and Managing Editor (Do not leave blank): Publisher—University of Illinois, Graduate School of Library and Information Science, Publications Office, 501 East Daniel Street, Champaign, Illinois 61820-6211; Editor—F. Wilfrid Lancaster, University of Illinois, Graduate School of Library and Information Science, 501 East Daniel Street, Champaign, Illinois 61820-6211; Managing Editor—James S. Dowling, University of Illinois, Graduate School of Library and Information Science, 501 East Daniel Street, Champaign, Illinois 61820-6211. (10) Owner: The Board of Trustees, University of Illinois, Urbana, Illinois 61801. (11) Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages, or Other Securities: None. (12) Tax Status: Has Not Changed During Preceding 12 Months. (13) Publication Title: *Library Trends.* (14) Issue Date for Circulation Data: 6/24/1999.

(15) Extent and Nature of Circulation	Average No. Copies Each Issue During Preceding 12 Months	No. Copies of Single Issue Published Nearest to Filing Date
a. Total Number of Copies (Net Press Run)	2858	2880
b. Paid and/or Requested Circulation		
(1) Paid Requested Outside-County Mail Subscriptions Stated on Form 3541.	2247	2306
(2) Paid in-County Subscriptions.	0	0
(3) Sales through Dealers and Carriers, Street Vendors, Counter Sales, and Other Non-USPS Paid Distribution.	0	0
(4) Other Classes Mailed Through the USPS.	0	0
c. Total Paid and/or Requested Circulation	2247	2306
d. Free Distribution by Mail		
(1) Outside-County as Stated on Form 3541	6	3
(2) In-County as Stated on Form 3541	0	0
(3) Other Classes Mailed Through the USPS	0	0
e. Free Distribution Outside the Mail	13	27
f. Total Free Distribution	19	30
g. Total Distribution	2266	2336
h. Copies not Distributed	592	544
i. Total	2858	2880
j. Percent Paid and/or Requested Circulation	99.16	98.72

(16) Publication of Statement of Ownership. Publication Required. Will be printed in the Volume 48, Number 2, Fall 1999 issue of this publication.